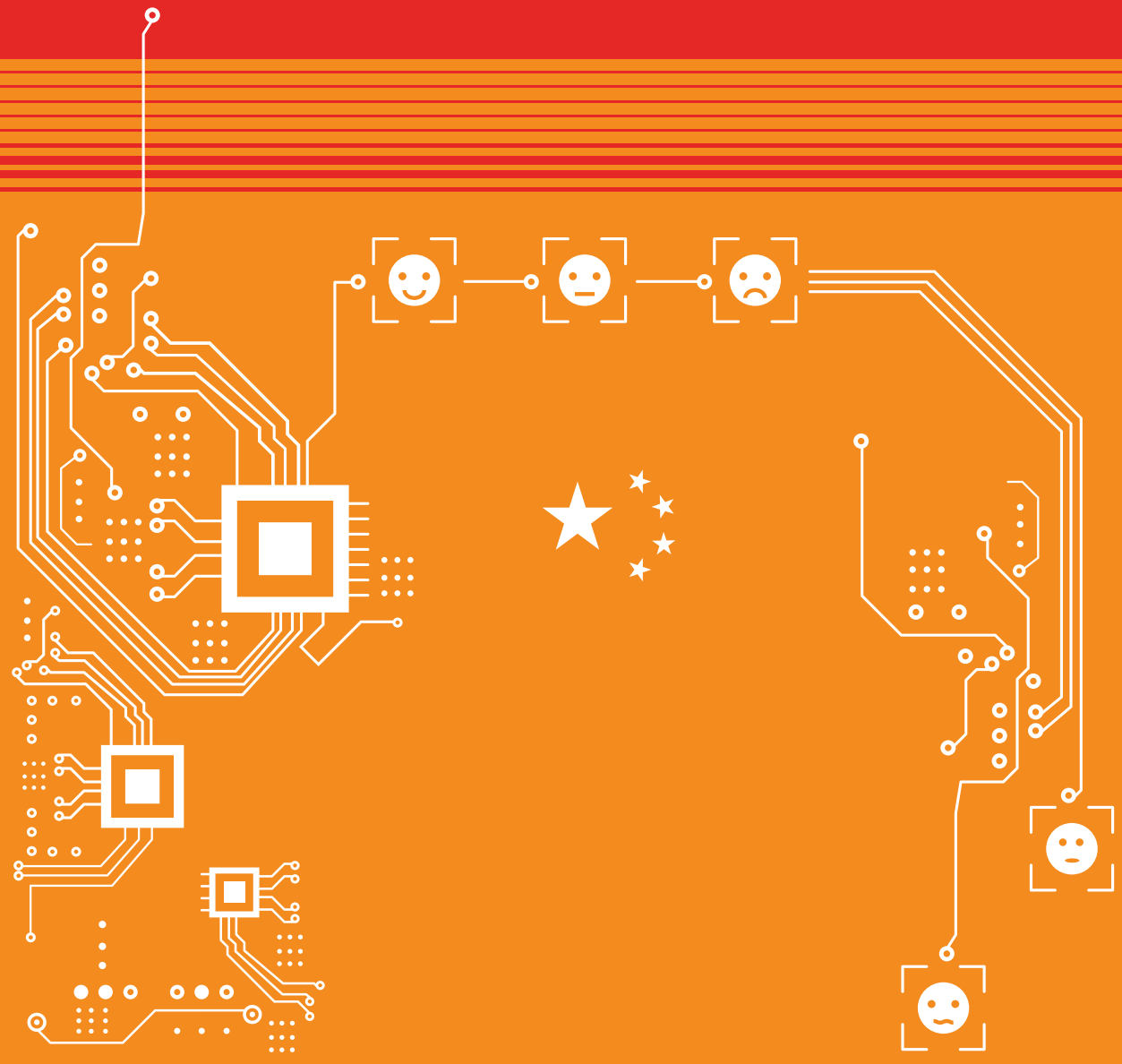
ARTICLE 19

Emotional Entanglement:

China's emotion recognition market and its implications for human rights



First published by ARTICLE 19 in January 2021

ARTICLE 19 works for a world where all people everywhere can freely express themselves and actively engage in public life without fear of discrimination. We do this by working on two interlocking freedoms, which set the foundation for all our work. The Freedom to Speak concerns everyone's right to express and disseminate opinions, ideas and information through any means, as well as to disagree from, and question power-holders. The Freedom to Know concerns the right to demand and receive information by power-holders for transparency good governance and sustainable development. When either of these freedoms comes under threat, by the failure of power-holders to adequately protect them, ARTICLE 19 speaks with one voice, through courts of law, through global and regional organisations, and through civil society wherever we are present.

ARTICLE 19

Free Word Centre

60 Farringdon Road

London EC1R 3GA

UK

www.article19.org

A19/DIG/2021/001

Contents

Executive Summary 5

Acknowledgements 9

Glossary 10

List of Abbreviations 11

Introduction 12

Why China? 13

Methodology 14

Background to Emotion Recognition 15

What Are Emotion Recognition Technologies? 15

How Reliable is Emotion Recognition? 15

Use Cases 17

Paving the Way for Emotion Recognition in China 18

Public Security 19

Foreign Emotion Recognition Precursors as Motivation 19

Three Types of Security-Use Contexts and Their Rationales 19

Public Security Implementations of Emotion Recognition 20

Driving Safety 23

In-Vehicle Emotion Recognition 23

Insurance Companies and Emotion Recognition of Drivers 23

Emotion Recognition Outside of Cars 24

State and Tech Industry Interest 24

Education 25

Emotion and Edtech 25

China's Push for 'AI+Education' 25

Chinese Academic Research on Emotion Recognition in Education 25

China's Market for Emotion Recognition in Education 26

Emotion Recognition in Online and In-Person Classrooms 29

Students' Experiences of Emotion Recognition Technologies 30

Parents' Perceptions of Emotion Recognition Technologies 34

Teachers' Experiences of Emotion Recognition Technologies 31

School Administrators' Perceptions of Emotion Recognition Technologies 32

Emotion Recognition and Human Rights 35
Right to Privacy 36
Right to Freedom of Expression 37
Right to Protest 38
Right Against Self-Incrimination 38
Non-Discrimination 38
Other Technical and Policy Considerations 39
Function Creep 39
Growing Chorus of Technical Concerns 39
Misaligned Stakeholder Incentives 40
Regional and Global Impact 40
Ethnicity and Emotion 40
Companies' Claims About Mental Health and Neurological Conditions 41
Emotion and Culpability 42
China's Legal Framework and Human Rights 44
China's National Legal Framework 45
Relationship to International Legal Frameworks 45
National Law 45
Chinese Constitution 45
Data Protection 45
Instruments 46
Biometric Data 47
Standardisation 47
Ethical Frameworks 48
Recommendations 49
To the Chinese Government 50
To the International Community 50
To the Private Companies Investigated in this Report 50
To Civil Society and Academia 50
Endnotes 51

Executive Summary

In this report, ARTICLE 19 provides evidence and analysis of the burgeoning market for emotion recognition technologies in China and its detrimental impact on individual freedoms and human rights, in particular the right to freedom of expression. Unlike better-known biometric applications, like facial recognition, that focus on identifying individuals, emotion recognition purports to infer a person's inner emotional state. Applications are increasingly integrated into critical aspects of everyday life: law enforcement authorities use the technology to identify 'suspicious' individuals, schools monitor students' attentiveness in class, and private companies determine people's access to credit.

Our report demonstrates the need for strategic and well-informed advocacy against the design, development, sale, and use of emotion recognition technologies. We emphasise that the timing of such advocacy – *before* these technologies become widespread – is crucial for the effective promotion and protection of people's rights, including their freedoms to express and opine. High school students should not fear the collection of data on their concentration levels and emotions in classrooms, just as suspects undergoing police interrogation must not have assessments of their emotional states used against them in an investigation. These are but a glimpse of uses for emotion recognition technologies being trialled in China.

This report describes how China's adoption of emotion recognition is unfolding within the country, and the prospects for the technology's export. It aims to:

1. Unpack and analyse the scientific foundations on which emotion recognition technologies are based;
2. Demonstrate the incompatibility between emotion recognition technology and international human rights standards, particularly freedom of expression, and the potential and ongoing detrimental impact of this technology on people's lives;
3. Provide rich detail on actors, incentives, and the nature of applications within three emotion recognition use cases in the Chinese market: public security, driving, and education;
4. Analyse the legal framework within which these use cases function; and
5. Set out recommendations for stakeholders, particularly civil society, on how to respond to the human rights threats posed by emotion recognition technologies in China.

This report will better equip readers to understand the precise ways in which China's legal, economic, and cultural context is different, the ways in which it is not, and why such distinctions matter. Each use case bears its own social norms, laws, and claims for how emotion recognition improves upon an existing process. Likewise, the interaction between pre-existing Chinese surveillance practices and these use cases shapes the contributions emotion recognition will make in China and beyond.

The implications of the report's findings are twofold. First, a number of problematic assumptions (many based on discredited science) abound amongst stakeholders interested in developing and/or deploying this technology. This report unpacks and critically analyses the human rights implications of emotion recognition technologies and the assumptions implicit in their marketing in China. Second, Chinese tech firms' growing influence in international technical standards-setting could encompass standards for emotion recognition. Using a human rights lens, the report addresses the most problematic views and practices that, if uncontested, could become codified in technical standards – and therefore reproduced in technology at a massive scale – at technical standard-setting bodies, like the International Telecommunications Union (ITU) and the Institute of Electrical and Electronics Engineers (IEEE).

Some of the main findings from the research on deployment of emotion recognition technology in China include the following:

The design, development, sale, and use of emotion recognition technologies are inconsistent with international human rights standards. While emotion recognition is fundamentally problematic, given its discriminatory and discredited scientific foundations, concerns are further exacerbated by how it is used to surveil, monitor, control access to opportunities, and impose power, making the use of emotion recognition technologies untenable under international human rights law (pp. 36–44).

The opaque and unfettered manner in which emotion recognition is being developed risks depriving people of their rights to freedom of expression, privacy, and the right to protest, amongst others. Our investigation reveals little evidence of oversight mechanisms or public consultation surrounding emotion recognition technologies in China, which contributes significantly to the speed and scale at which use cases are evolving. Mainstream media is yet to capture the nuance and scale of this burgeoning market, and evidence collection is crucial at this moment. Together, these factors impede civil society's ability to advocate against this technology.

Emotion recognition's pseudoscientific foundations render this technology untenable as documented in this report. Even as some stakeholders claim that this technology can get better with time, given the pseudoscientific and racist foundations of emotion recognition on one hand, and fundamental incompatibility with human rights on the other, the design, development, deployment, sale, and transfer of these technologies must be banned.

Emotion recognition technologies' flawed and long-discredited scientific assumptions do not hinder their market growth in China. Three erroneous assumptions underlie justifications for the use and sale of emotion recognition technologies: that facial expressions are universal, that emotional states can be unearthed from them, and that such inferences are reliable enough to be used to make decisions. Scientists across the world have discredited all three assumptions for decades, but this does not seem to hinder the experimentation and sale of emotion recognition technologies (pp. 18–35).

Chinese law enforcement and public security bureaux are attracted to using emotion recognition software as an interrogative and investigatory tool. Some companies seek procurement order contracts for state surveillance projects (pp. 18-22) and train police to use their products (p. 22). Other companies appeal to law enforcement by insinuating that their technology helps circumvent legal protections concerning self-incrimination for suspected criminals (pp. 42-43).

While some emotion recognition companies allege they can detect sensitive attributes, such as mental health conditions and race, none have addressed the potentially discriminatory consequences of collecting this information in conjunction with emotion data. Some companies' application programming interfaces (APIs) include questionable racial categories for undisclosed reasons (p. 41). Firms that purportedly identify neurological diseases and psychological disorders from facial emotions (pp. 41-42) fail to account for how their commercial emotion recognition applications might factor in these considerations when assessing people's emotions in non-medical settings, like classrooms.

Chinese emotion recognition companies' stances on the relationship between cultural background and expressions of emotion influence their products. This can lead to problematic claims about emotions being presented in the same way across different cultures (p. 40) – or, conversely, to calls for models trained on 'Chinese faces' (p. 41). The belief that cultural differences do not matter could result in inaccurate judgements about people from cultural backgrounds that are underrepresented in the training data of these technologies – a particularly worrying outcome for ethnic minorities.

Chinese local governments' budding interest in emotion recognition applications confer advantages to both startups and established tech firms. Law enforcement institutions' willingness to share their data with companies for algorithm-performance improvement (p. 22), along with local government policy incentives (pp. 18, 20, 22, 24, 25, 33), enable the rapid development and implementation of emotion recognition technologies.

The emotion recognition market is championed by not only technology companies but also partnerships linking academia, tech firms, and the state. Assertions about emotion recognition methods and applications travel from academic research papers to companies' marketing materials (pp. 22, 25-26) and to the tech companies' and state's public justifications for use (pp. 20, 22-33). These interactions work in tandem to legitimise uses of emotion recognition that have the potential to violate human rights.

None of the Chinese companies researched here appears to have immediate plans to export their products. Current interest in export seems low, (p. 40) although companies that already have major markets abroad, such as Hikvision and Huawei, are working on emotion recognition applications (pp. 23, 27, 29-33, 40).

People targeted by these technologies in China – particularly young adults (pp. 30–31**) – predominantly report feeling distrust, anxiety, and indifference regarding current emotion recognition applications in education.** While some have criticised emotion recognition in education-use scenarios (pp. 30-31, 34), it is unclear whether there will be ongoing pushback as awareness spreads.

Civil society strategies for effective pushback will need to be tailored to the context of advocacy. Civil society interventions can focus on debunking emotion recognition technology's scientific foundations, demonstrating the futility of using it, and/or demonstrating its incompatibility with human rights. The strategy (or strategies) that civil society actors eventually employ may need to be adopted in an agile manner that considers the geographic, political, social, and cultural context of use.

Acknowledgements

ARTICLE 19 is grateful to Graham Webster, Jeffrey Ding, Luke Stark, and participants at the RealML 2020 workshop for their insightful feedback on various drafts of this report.

If you would like to discuss any aspects of this report further, please email info@article19.org to get in touch with:

1. Vidushi Marda, Senior Programme Officer, ARTICLE 19
2. Shazeda Ahmed, PhD candidate, UC Berkeley School of Information

Glossary

Biometric data: Data relating to physical, physiological, or behavioural characteristics of a natural person, from which identification templates of that natural person – such as faceprints or voice prints – can be extracted. Fingerprints have the longest legacy of use for forensics and identification,[1] while more recent sources include (but are not limited to) face, voice, retina and iris patterns, and gait.

Emotion recognition: A biometric application that uses machine learning in an attempt to identify individuals' emotional states and sort them into discrete categories, such as anger, surprise, fear, happiness, etc. Input data can include individuals' faces, body movements, vocal tone, spoken or typed words, and physiological signals (e.g. heart rate, blood pressure, breathing rate).

Facial recognition: A biometric application that uses machine learning to identify (1:n matching) or verify (1:1 matching) individuals' identities using their faces. Facial recognition can be done in real time or asynchronously.

Machine learning: A popular technique in the field of artificial intelligence that has gained prominence in recent years. It uses algorithms trained with vast amounts of data to improve a system's performance at a task over time.

Physiognomy: The pseudoscientific practice of using people's outer appearance, particularly the face, to infer qualities about their inner character.

List of Abbreviations

AI	Artificial intelligence
BET	Basic Emotion Theory
CCS	Class Care System
DRVR	Driving Risk Video Recognition
FACE KYD	Face Know Your Driver
GDPR	General Data Protection Regulation
HRC	UN Human Rights Council
ICCPR	International Covenant on Civil and Political Rights
ICT	Information and communications technologies
ITU	International Telecommunications Union
MOOC	Massive open online courses
OBOR	One Belt, One Road
PSB	Public security bureau
SPOT	Screening of Passengers by Observation Techniques
TAL	Tomorrow Advancing Life
UE	Universal facial expressions

1. Introduction

Biometric technologies, particularly face-based biometric technologies, are increasingly used by states and private actors to identify, authenticate, classify, and track individuals across a range of contexts – from public administration and digital payments to remote workforce management – often without their consent or knowledge.[2] States have also been using biometric technologies to identify and track people of colour, suppress dissent, and carry out wrongful arrests, even as a rapidly growing body of research has demonstrated that these systems perform poorly on the faces of Black women, ethnic minorities, trans people, and children.[3]

Human rights organisations, including ARTICLE 19, have argued that public and private actors' use of biometrics poses profound challenges for individuals in their daily lives, from wrongfully denying welfare benefits to surveilling and tracking vulnerable individuals with no justifications. As they are currently used, biometric technologies thus pose disproportionate risks to human rights, in particular to individuals' freedom of expression, privacy, freedom of assembly, non-discrimination, and due process. A central challenge for civil society actors and policymakers thus far is that pushback against these technologies is often reactive rather than proactive, reaching a crescendo only after the technologies have become ubiquitous.[4]

In an attempt to encourage pre-emptive and strategic advocacy in this realm, this report focuses on emotion recognition, a relatively under-observed application of biometric technology, which is slowly entering both public and private spheres of life. Emerging from the field of affective computing,[5] emotion recognition is projected to be a USD65 billion industry by 2024,[6] and is already cropping up around the world.[7] Unlike any ubiquitous biometric technology, it claims to infer individuals' inner feelings and emotional states, and a ground truth about a subjective, context-dependent state of being. While face recognition asked who we *are*, emotion recognition is chiefly concerned with how we *feel*. Many believe this is not possible to prove or disprove.[8]

In this report, ARTICLE 19 documents the development, marketing, and deployment of emotion recognition in China, and examines the various actors, institutions, and incentives that bring these technologies into existence.

We discuss the use of emotion recognition in three distinct sectors in China: public security, driving safety, and education. In doing so, the report foregrounds how civil society will face different sets of social norms, policy priorities, and assumptions about how emotion recognition serves each of these three sectors. At the same time, these sectors share some commonalities:

1. They all hint at how 'smart city' marketing will encompass emotion recognition.

2. They all take place in spaces that people often have no choice in interacting with, leaving no substantial consent or opt-out mechanisms for those who do not want to participate.

3. Although major Chinese tech companies – including Baidu and Alibaba – are experimenting with emotion recognition, this report focuses on the majority of commercial actors in the field: smaller startups that go unnoticed in major English-language media outlets, but that have nonetheless managed to link up with academics and local governments to develop and implement emotion recognition.

Why China?

This report focuses on China because it is a dominant market with the technologically skilled workforce, abundant capital, market demand, political motivations, and export potential for artificial intelligence (AI) that could enable rapid diffusion of emotion recognition technologies.[9] Over the past few years, Chinese tech companies have fuelled an international boom in foreign governments' acquisition of surveillance technology.[10] China's One Belt, One Road (OBOR) initiative has enabled the wide-scale

implementation of Huawei's Safe Cities policing platforms and Hikvision facial recognition cameras, in democracies and autocracies alike, without accompanying public deliberation or safeguards. In the context of facial recognition in particular, policymakers were taken aback by how quickly the Chinese companies that developed this technology domestically grew and started to export their products to other countries.[11]

Discussing emotion recognition technologies, Rosalind Picard – founder of major affective computing firm, Affectiva, and one of the leading researchers in the field – recently commented:

> *"The way that some of this technology is being used in places like China, right now […] worries me so deeply, that it's causing me to pull back myself on a lot of the things that we could be doing, and try to get the community to think a little bit more about [...] if we're going to go forward with that, how can we do it in a way that puts forward safeguards that protect people?"*[12]

To effectively advocate against emotion recognition technologies, it is crucial to concentrate on the motivations and incentives of those Chinese companies that are proactive in proposing international technical standards for AI applications, including facial recognition, at convening bodies like the ITU.[13] Internationally, a head start on technical standards-setting could enable Chinese tech companies to develop interoperable systems and pool data, grow more globally competitive, lead international governance on AI safety and ethics, and obtain the 'right to speak' that Chinese representatives felt they lacked when technical standards for the Internet were set.[14] This codification reverberates throughout future markets for this particular technology, expanding the technical standards' worldwide influence over time.

Focusing on the Chinese emotion recognition *market*, in particular, provides an opportunity to pre-empt how China's embrace of emotion recognition can – and will – unfold outside of China's borders. If international demand for emotion recognition increases, China's pre-existing market for technology exports positions a handful of its companies to become major suppliers, following on the heels of their dominance of the facial recognition market.[15] With this report, ARTICLE 19 therefore seeks to galvanise civil society attention to the increasing use of emotion recognition technologies, their pseudoscientific underpinnings, and the fundamental inconsistency of their commercial applications with international human rights standards. We seek to do so early in emotion recognition's commercialisation, before it is widespread globally, to pre-empt the blunt and myopic ways in which adoption of this technology might grow.

Methodology

The research for this report began with a literature review built from Mandarin-language sources in two Chinese academic databases: China National Knowledge Infrastructure and the Superstar Database (超星期刊). Search keywords included terms related to emotion recognition (情绪识别), micro-expression recognition (微表情识别), and affective computing (情感计算). In parallel, the authors consulted Chinese tech company directory Tianyancha (天眼查), where 19 Chinese companies were tagged as working on emotion recognition. Of these, eight were selected for further research because they provided technology that fit within the three use cases the report covers. The additional 19 companies investigated came up in academic and news media articles that mentioned the eight firms chosen from the Tianyancha set, and were added into the research process. Google, Baidu, and WeChat Mandarin-language news searches for these companies, as well as for startups and initiatives unearthed in the academic literature, formed the next stage of source collection.

Finally, where relevant, the authors guided a research assistant to find English-language news and academic research that shed light on comparative examples.

We mention and analyse these 27 companies based on the credibility and availability of source material, both within and outside company websites, and examples of named institutions that have pilot tested or fully incorporated these companies' products. For a few companies, such as Miaodong

in Guizhou, news coverage is not recent and it is unclear whether the company is still operating. Nonetheless, such examples were included alongside more recently updated ones to highlight details that are valuable to understanding the broader trend of emotion recognition applications, such as access to law enforcement data for training emotion recognition models, or instances where public pushback led to modification or removal of a technology. Even if some of these companies are defunct, a future crop of competitors is likely to follow in their stead.

Finally, although other types of emotion recognition that do not rely on face data are being used in China, the report focuses primarily on facial expression-based and multimodal emotion recognition that includes face analysis, as our research revealed these two types of emotion recognition are more likely to be used in high-stakes settings.

Background to Emotion Recognition

What Are Emotion Recognition Technologies?

Emotion recognition technologies purport to infer an individual's inner affective state based on traits such as facial muscle movements, vocal tone, body movements, and other biometric signals. They use machine learning (the most popular technique in the field of AI) to analyse facial expressions and other biometric data and subsequently infer a person's emotional state.[16]

Much like other biometric technologies (like facial recognition), the use of emotion recognition involves the mass collection of sensitive personal data in invisible and unaccountable ways, enabling the tracking, monitoring, and profiling of individuals, often in real time.[17]

Some Chinese companies describe the link between facial recognition technologies (based on comparing faces to determine a match) and emotion recognition (analysing faces and assigning emotional categories to them) as a matter of incremental progress. For example, Alpha Hawkeye (阿尔法鹰眼), a Chinese company that supplies emotion recognition for public security, characterises it as 'biometrics 3.0'[18], while a write-up of another company, Xinktech (云思创智), predicts 'the rise of emotion recognition will be faster than the face recognition boom, because now there is sufficient computing power and supporting data. The road to emotion recognition will not be as long.'[19]

How Reliable is Emotion Recognition?

Two fundamental assumptions undergird emotion recognition technologies: that it is possible to gauge a person's inner emotions from their external expressions, and that such inner emotions are both discrete and uniformly expressed across the world. This idea, known as Basic Emotion Theory (BET), draws from psychologist Paul Ekman's work from the 1960s. Ekman suggested humans across cultures could reliably discern emotional states from facial expressions, which he claimed were universal.[20] Ekman and Friesen also argued that micro-momentary expressions ('micro-expressions'), or facial expressions that occur briefly in response to stimuli, are signs of 'involuntary emotional leakage [which] exposes a person's true emotions'.[21]

BET has been wildly influential, even inspiring popular television shows and films.[22] However, scientists have investigated, contested, and largely rejected the validity of these claims since the time of their publication.[23] In a literature review of 1,000 papers' worth of evidence exploring the link between emotional states and expressions, a panel of authors concluded:

> *"very little is known about how and why certain facial movements express instances of emotion, particularly at a level of detail sufficient for such conclusions to be used in important, real-world applications. Efforts to simply 'read out' people's internal states from an analysis of their facial movements alone, without considering various aspects of context, are at best incomplete and at worst entirely lack validity, no matter how sophisticated the computational algorithms".*[24]

Another empirical study sought to find out whether the assumption that facial expressions are a consequence of emotions was valid, and concluded that 'the reported meta-analyses for happiness/amusement (when combined), surprise, disgust, sadness, anger, and fear found that all six emotions were on average only weakly associated with the facial expressions that have been posited as their UEs [universal facial expressions]'.[25]

The universality of emotional expressions has also been discredited through the years. For one, researchers found that Ekman's methodology to determine universal emotions inadvertently primed subjects (insinuated the 'correct' answers) and eventually distorted results.[26] The 'natural kind' view of emotions as something nature has endowed humans with, independent of our perception of emotions and their cultural context, has been strongly refuted as a concept that has 'outlived its scientific value and now presents a major obstacle to understanding what emotions are and how they work'.[27]

Finally, empirical studies have disproved the notion of micro-expressions as reliable indicators of emotions; instead finding them to be both unreliable (due to brevity and infrequency) and discriminatory.[28] Some scholars have proposed a 'minimum universality' of emotions, insisting 'the finite number of ways that facial muscles can move creates a basic template of expressions that are then filtered through culture to gain meaning'.[29] This is corroborated by a recent study from the University of Glasgow, which found that culture shapes the perception of emotions.[30] Yet even theories of minimum universality call the utility of AI-driven emotion recognition systems into question. One scholar has suggested that, even if such technologies 'are able to map each and every human face perfectly, the technical capacities of physiological classification will still be subject to the vagaries of embedded cultural histories and contemporary forms of discrimination and of racial ordering'.[31]

Even so, academic studies and real-world applications continue to be built on the basic assumptions about emotional expression discussed above, despite these assumptions being rooted in dubious scientific studies and a longer history of discredited and racist pseudoscience.[32]

Emotion recognition's application to identify, surveil, track, and classify individuals across a variety of sectors is thus **doubly problematic** – not just because of its dangerous applications, but also because it *doesn't even work* as its developers and users claim.[33]

2. Use Cases

Paving the Way for Emotion Recognition in China

As one of the world's biggest adopters of facial recognition cameras, China has come under scrutiny for its tech firms' far-reaching international sale of surveillance technology.[34] The normalisation of surveillance in Chinese cities has developed in parallel with the government's crackdown on the ethnic minority Uighur population in Xinjiang province. For Xinjiang's majority-Muslim population, security cameras, frequent police inspections, law enforcement's creation of Uighur DNA and voiceprint databases, and pervasive Internet monitoring and censorship of content about or related to Islam are inescapable.[35]

One state-sponsored security venture, the 'Sharp Eyes' project (雪亮工程), has come up in relation to three of the ten companies investigated in this section. Sharp Eyes is a nationwide effort to blanket Chinese cities and villages with surveillance cameras, including those with licence plate-reading and facial recognition capabilities.[36] The project, which the Central Committee of the Chinese Communist Party approved in 2016, relies in part on the government procurement-order bidding process to allocate billions of yuan in funding to (foreign and domestic) firms that build and operate this infrastructure.[37]

A homologous concept resurgent in contemporary surveillance is the 'Fengqiao experience' (枫桥经验), a Mao Zedong-contrived practice in which ordinary Chinese citizens monitored and reported each other's improper behaviour to the authorities. In a story that has come to exemplify Fengqiao, rock musician Chen Yufan was arrested for drug charges when a 'community tip' from within his residential area made its way to authorities.[38] President Xi Jinping has praised the return of the Fengqiao experience through neighbourhood-level community watch groups that report on suspected illegal behaviour. Though senior citizens are the backbone of this analogue surveillance, police have begun to head up watch groups, and technology companies have capitalised on the Fengqiao trend by developing local apps incentivising people to report suspicious activity in exchange for rewards, such as discounted products and services from major tech firms. In late 2018, a conference on digital innovation and social management, The New Fengqiao Experience, convened police officers and companies including Alibaba.[39]

Although reporting on Sharp Eyes and Fengqiao-style policing has not yet touched on emotion recognition, both are relevant for three reasons. For one, Sharp Eyes and the Fengqiao project exemplify templates for how multiple national government organisations, tech companies, and local law enforcement unite to implement surveillance technology at scale. Second, companies specialising in emotion recognition have begun to either supply technology to these projects or to incorporate both Sharp Eyes and Fengqiao into their marketing, as seen below with companies Alpha Hawkeye (阿尔法鹰眼), ZNV Liwei, and Xinktech.[40] Finally, Chinese tech firms' commercial framing of emotion recognition as a natural next step in the evolution of biometric technology applications opens up the possibility that emotion recognition will be integrated in places where facial recognition has been widely implemented. Independent researchers are already using cameras with image resolution sufficiently high to conduct face recognition in experiments to develop emotion and gesture recognition.[41]

It is important to note that interest in multimodal emotion recognition is already high. Media coverage of the company Xinktech predicts that micro-expression recognition will become a ubiquitous form of data collection, fuelling the rise of 'multimodal technology [as an] inevitable trend, a sharp weapon, and a core competitive advantage in the development of AI'.[42] By one estimate, the potential market for multimodal emotion recognition technologies is near 100 billion yuan (over USD14.6 billion).[43] How did multimodality garner such hype this early in China's commercial development of emotion recognition? Part of the answer lies in how Chinese tech firms depict foreign examples of emotion recognition as having been unilateral successes – ignoring the scepticism that terminated some of these initiatives.

Public Security

Foreign Emotion Recognition Precursors as Motivation

A popular theme in China's academic and tech industry literature about using emotion recognition for public security is the argument that it has achieved desirable results abroad. Examples cited include both automated and non-technological methods of training border-patrol and police officers to recognise micro-expressions, such as the US Transportation Security Authority's Screening Passengers by Observation Techniques (SPOT) programme and Europe's iBorderCtrl. Launched in 2007, SPOT was a programme that trained law enforcement officials known as Behaviour Detection officers to visually identify suspicious behaviours and facial expressions from the Facial Action Coding System. Chinese police academies' research papers have also made references to US plainclothes police officers similarly using human-conducted micro-expression recognition to identify terrorists – a practice Wenzhou customs officials dubbed 'worth drawing lessons from in our travel inspection work'.[44] iBorderCtrl, a short-lived automated equivalent trialled in Hungary, Latvia, and Greece, was a pre-screening AI system whose cameras scanned travellers' faces for signs of deception while they responded to border-security agents' questions.

A major omission in the effort to build a case for emotion recognition in Chinese public security is that much of what passes for 'success' stories has been derided for instances that have been heavily contested and subject of legal challenge for violation of human rights. The American Civil Liberties Union, Government Accountability Office, Department of Homeland Security, and even a former SPOT officer manager have exposed the SPOT programme's unscientific basis and the racial profiling it espoused.[45] Officers working on this programme told the *New York Times* that they "just pull aside anyone who they don't like the way they look — if they are Black and have expensive clothes or jewellery, or if they are Hispanic".[46] iBorderCtrl's dataset has been criticised for false positives, and its discriminatory potential led to its retraction.[47]

When discussed in Chinese research, news, and marketing, these final outcomes are glossed over – such as in a feature on Alpha Hawkeye, which made the unsourced claim that the SPOT programme's cost per individual screening was USD20, in comparison to Alpha Hawkeye's USD0.80 per inspection.[48]

Three Types of Security-Use Contexts and Their Rationales

Emotion recognition software and hardware that are implemented in security settings fall into three categories:

1. 'Early warning' (预警);[49]
2. Closer monitoring after initial identification of a potential threat; and
3. Interrogation.

The firms' marketing approaches vary depending on the category of use. Sometimes marketed as more scientific, accurate descendants of lie-detection (polygraph) machines, emotion recognition-powered interrogation systems tend to extract facial expressions, body movements, and vocal tone from video recordings. In particular, the academic literature coming out of police-training academies provides the boilerplate justifications that tech companies reproduce in their marketing materials.

One Chinese research paper from the Hubei Police Academy discusses the value of facial micro-expressions in identifying 'dangerous people' and 'high-risk groups' who do not have prior criminal records.[50] The author proposes creating databases that contain video images of criminals before and after they have committed crimes, as a basis for training algorithms that can pick up on the same facial muscle movements and behaviours in other people.[51] The argument driving this – and all uses of emotion recognition in public security settings – is the belief that people feel guilt before committing a crime, and that they cannot mask this 'true' inner state in facial expressions so minor or fleeting that only high-resolution cameras can detect them.[52]

Another paper from two researchers at Sichuan Police College envisioned a Tibetan border-patrol inspection system that would fit both the 'early warning' and follow-up inspection functions.[53] They argued that traditional border-security inspections can be invasive and time-consuming, and that the longer they take, the more the individuals being inspected feel they are being discriminated against.[54] Yet if AI could be used to identify suspicious micro-expressions, they reasoned, presumably fewer people would be flagged for additional inspection, and the process would be less labour-intensive for security personnel. Moreover, the speed of the automated process is itself presented as somehow 'fairer' for those under inspection by taking up less of their time. In a similar framing to the Hubei Police Academy paper, the authors believed their system would be able to root out 'Tibetan independence elements' on the basis of emotion recognition.[55] These disconcerting logical leaps are replicated in how the companies themselves market their products.

Public Security Implementations of Emotion Recognition

News coverage and marketing materials for the ten companies described in Table 1 flesh out the context in which emotion recognition applications are developed.

According to one local news story, authorities at the Yiwu Railway Station (Zhejiang) used Alpha Hawkeye's emotion recognition system to apprehend 153 so-called 'criminals' between October 2014 and October 2015.[56] The headline focused on the more mundane transgression that these types of systems tend to over-police: individuals' possession of two different state ID cards. Alpha Hawkeye's products have reportedly been used in both Sharp Eyes projects and in the OBOR 'counterterrorism industry'.[57] ZNV Liwei (ZNV力维) is also reported to have contributed technology to the Sharp Eyes surveillance project and to have provided police in Ningxia, Chongqing, Shenzhen, Shanghai, and Xinjiang with other 'smart public security products', though the company's website does not indicate whether emotion recognition capabilities are among them.[58] An article from 2017 indicated that Alpha Hawkeye planned to develop its own 'high-risk crowd database' that would match footage collected from its cameras against (unnamed) 'national face recognition databases'.[59] In coordination with local authorities, the company has conducted pilot tests in rail and subway stations in Beijing, Hangzhou, Yiwu (Zhejiang), Urumqi (Xinjiang), and Erenhot (Inner Mongolia), at airports in Beijing and Guangzhou, and at undisclosed sites in Qingdao and Jinan, although it is ambiguous about whether these applications involved only face recognition or also included emotion recognition.[60]

The user interface for an interrogation platform from CM Cross (深圳市科思创动科技有限公司, known as 科思创动) contains a 'Tension Index Table' (紧张程度指数表) that conveys the level of tension a person under observation supposedly exhibits, with outputs including 'normal', 'moderate attention', and 'additional inspection suggested'.[61] Moreover, the CM Cross interrogation platform sorts questions to pose to suspects into interview types; for example, 'conventional interrogations', 'non-targeted interviews', and 'comprehensive cognitive tests'.[62]

At the 8th China (Beijing) International Police Equipment and Counter-Terrorism Technology Expo in 2019, Taigusys Computing representatives marketed their interrogation tools as obviating the need for polygraph machines, and boasted that their prison-surveillance system can prevent inmate self-harm and violence from breaking out by sending notifications about inmates expressing 'abnormal emotions' to on-site management staff. Images of the user interface for the 'Mental Auxiliary Judgment System' (精神辅助判定系统) on the company's website show that numerical values are assigned to nebulous indicators, such as 'physical and mental balance' (身心平衡).[63]

Table 1: Companies Providing Emotion Recognition for Public Security

Company Name	Products and Methods of Data Collection	Suggested Uses[64]
Alpha Hawkeye 阿尔法鹰眼	Monitors vestibular emotional reflex and conducts posture, speech, physiological, and semantic analysis.[65]	• Airport, railway, and subway station early-warning threat detection • Customs and border patrol
CM Cross 科思创动	Employs deep-learning-powered image recognition to detect blood pressure, heart rate, and other physiological data.[66]	• Customs and border patrol[67] • Early warning • Police and judicial interrogations
EmoKit 翼开科技	EmoAsk AI Multimodal Smart Interrogation Auxiliary System detects facial expressions, body movements, vocal tone, and heart rate.[68] Other products detect similar data for non-interrogation uses.	• Detecting and managing mental-health issues at medical institutions • Loan interviews at banks • Police-conducted interrogations[69] and other law enforcement-led questioning of convicted criminals[70]
Joyware 中威电子 NuraLogix	NuraLogix's DeepAffex is an image recognition engine that identifies facial blood flow (which is used to measure emotions) and detects heart rate, breathing rate, and 'psychological pressure'.[71] Joyware also uses NuraLogix's polygraph tests.[72]	• Airport and railway station surveillance • Nursing • Psychological counselling
Miaodong 秒懂	Relies on image recognition of vibrations and frequency of light on faces, which are used to detect facial blood flow and heart rate as a basis for emotion recognition.[73]	• Police interrogation
Sage Data 睿数科技	Public Safety Multimodal Emotional Interrogation System detects micro-expressions, bodily micro-actions, heart rate, and body temperature.[74]	• Police and court interrogations
Shenzhen Anshibao 深圳安视宝	Emotion recognition product detects frequency and amplitude of light vibrations on faces and bodies, which Shenzhen Anshibao believes can be used to detect mental state and aggression.[75]	• Early warning[76] • Prevention of crimes and acts of terror
Taigusys Computing 太古计算	One product is referred to as a micro-expression-recognition system for Monitoring and Analysis of Imperceptible Emotions at Interrogation Sites, while others include 'smart prison' and 'dynamic emotion recognition' solutions. Taigusys claims to use image recognition that detects light vibrations on faces and bodies, as well as parallel computing.[77]	• Hospital use for detecting Alzheimer's, depression, and panic attacks[78] • Police interrogation of suspected criminals[79] • Prison surveillance
Xinktech 云思创智	Products include 'Lingshi' Multimodal Emotional Interrogation System and Public Security Multimodal Emotion Research and Judgment System, among others. They can detect eight emotions and analyses facial expression, posture, semantic, and physiological data.[80]	• Judicial interrogation[81] • Police interrogation[82] • Public security settings, including customs inspections[83]
ZNV Liwei ZNV力维	Collects data on heart rate and blood-oxygen level.[84]	• Police interrogation of suspected criminals

Xinktech (南京云思创智科技公司) aims to create the 'AlphaGo of interrogation'.[85] Their 'Lingshi' Multimodal Emotional Interrogation System' (灵视多模态情绪审讯系统), showcased at the Liupanshui 2018 criminal defence law conference in Hubei, contains 'core algorithms that extract 68 facial feature points and can detect eight emotions (calmness, happiness, sadness, anger, surprise, fear, contempt, disgust).[86] Aside from providing a venue for the companies to showcase their products, conferences double as a site for recruiting both state and industry partners in development and implementation.

In 2018, Hangzhou-based video surveillance firm Joyware signed a cooperative agreement to develop 'emotional AI' with the Canadian image recognition company NuraLogix.[87] NuraLogix trains models to identify facial blood flow as a measure of emotional state and other vital signs.[88] ZNV Liwei has collaborated with Nanjing Forest Police College and CM Cross to establish an 'AI Emotion Big Data Joint Laboratory' (AI情绪大数据联合实验室), where they jointly develop 'psychological and emotion recognition big data systems'.[89] In 2019, Xinktech held an emotion recognition technology seminar in Nanjing. Media coverage of the event spotlighted the company's cooperative relationship with the Interrogation Science and Technology Research Center of the People's Public Security University of China, along with Xinktech's joint laboratory with the Institute of Criminal Justice at Zhongnan University of Economics and Law established earlier that year.[90]

Xinktech's partnerships with both of these universities and Nanjing Forest Police Academy account for some of its training data acquisition and model-building process – contributions that reflect a symbiotic exchange between firms and the state.[91] EmoKit (翼开科技), which professed to have 20 million users of its open APIs four years ago, partnered with the Qujing Public Security Bureau (PSB) in Yunnan Province.[92] According to one source, EmoKit obtained 20 terabytes of interrogation video data from a southern Chinese police department.[93] In Guizhou, a startup called Miaodong (秒懂) received a similar boost from local government in 2016.[94] At first, Miaodong's interrogation software was reputedly only accurate 50% of the time. They then came to the attention of local officials in the Guiyang High Tech Zone and teamed up with the Liupanshui PSB. After this, the PSB shared several archived police interrogation videos with Miaodong, and the company says its accuracy rates rose to 80%.[95] Similarly, Xinktech partnered with police officers to label over 2,000 hours of video footage containing 4 million samples of emotion image data. When asked why Xinktech entered the public security market, CEO Ling responded: "We discovered that the majority of unicorns in the AI field are companies who start out working on government business, mainly because the government has pain points, funding, and data."[96] Exploiting these perceived 'pain points' further, some companies offer technology training sessions to law enforcement.

At a conference, Xinktech CEO Ling Zhihui discussed the results of Xinktech's product applications in Wuxi, Wuhan, and Xinjiang.[97] Afterwards, Ling facilitated a visit to the Caidian District PSB in Wuhan to demonstrate their pilot programme using Xinktech's 'Public Security Multimodal Emotion Research and Judgment System' (公安多模态情绪研判系统).[98] Xinktech reportedly also sells its 'Lingshi' interrogation platform to public security and prosecutorial institutions in Beijing, Hebei, Hubei, Jiangsu, Shaanxi, Shandong, and Xinjiang.[99] Concurrently with the Hubei conference, Xinktech's senior product manager led the 'Interrogation Professionals Training for the Province-Wide Criminal Investigation Department' (全省刑侦部门审讯专业人才培训) at the Changzhou People's Police Academy in Jiangsu province, an event co-sponsored by the Jiangsu Province Public Security Department.[100] Finally, in late 2019, EmoKit's CEO described a pilot test wherein police in Qujing, Yunnan, would trial the company's interrogation technology. EmoKit planned to submit results from this test run in its application to join the list of police equipment procurement entities that supply the Ministry of Public Security.[101] EmoKit also purports to work with the military, with one military-cooperation contract raking in 10 million RMB (USD1.5 million USD), compared with 1 million RMB (USD152,000 USD) orders in the financial and education sectors, respectively.[102]

Driving Safety

The span of driving-safety applications of emotion recognition runs from in-car interventions to stationary hardware mounted on roadways. As with the other ucse cases in this report, this subsector of applications is not unique to China.[103] All of the Chinese examples in this section feature emotion sensing, in addition to driver-fatigue detection, and notably seem to group both under emotion or expression recognition.

In-Vehicle Emotion Recognition

Smart car manufacturer LeEco was reported to have incorporated face and emotion recognition into its LeSee concept car model in 2016.[104] In its 2019 corporate social responsibility report, Great Wall Motors announced that in at least three of its models it had launched an 'intelligent safety system', Collie, which includes 'emotion/expression recognition' and facial recognition capabilities among a total of 43 features to protect drivers, passengers, and pedestrians.[105] A reporter who tested one of these Great Wall Motors models, the VV7, found that when the car's emotion recognition technology sensed the reporter was 'angry' it automatically played more up-tempo music.[106] Additional media coverage of Great Wall Motor's VV6 model, which is reported to be released in 2021, indicates that the emotion recognition system can be continually upgraded as firmware-over-the-air, such that the emotion and fatigue recognition system can receive push updates of 'relevant' music.[107]

When state-owned car manufacturer Chang'an Automobiles promoted its UNI-T SUV crossover model at a connected-car technology expo in April 2020, media coverage described the in-vehicle UNI-T system as able to detect drivers' emotions and fatigue levels through facial emotion recognition.[108] Frequent yawning and blinking might prompt the UNI-T system to verbally warn the driver to be more alert, or – as with the Great Wall Motors cars – the system might automatically play 'rejuvenating' music.[109]

Aside from automobile manufacturers, hardware companies and AI startups are also contributing to the emerging trend of outfitting cars with emotion recognition functions. For instance, in late 2020, Huawei showcased its HiCar system that links drivers' mobile phones to their cars, enabling applications of computer vision, including emotion recognition and driver-fatigue recognition.[110] Taigusys Computing, the company that has provided emotion and behaviour recognition cameras for monitoring prisons and schools, has likewise developed a 'driver abnormal behaviour recognition system' that assesses drivers' facial expressions, body movements, and the content of their speech to issue early warnings if any of these actions is deemed unsafe.[111]

While most instances of in-vehicle emotion recognition focus on drivers, one Chinese car manufacturer has chosen to broaden its scope to additionally identify the emotional states of passengers. AIWAYS (爱驰汽车) has developed 'smart companion technology' that news reports describe as being able to detect a child passenger's emotions that may distract a parent's driving. If a child is crying in the backseat, the AIWAYS system can 'appease the child by playing songs the child likes, stories, and even sounds of the child's own happy laughter'.[112]

Insurance Companies and Emotion Recognition of Drivers

Insurance providers have also begun turning to emotion recognition to streamline their operations. China's biggest insurance firm, Ping An Group, demonstrated an in-vehicle facial expression recognition system that merges two of the company's products, Face Know Your Driver (FACE KYD) and Driving Risk Video Recognition (DRVR), at an expo in late 2019. The former extracts drivers' facial micro-expressions in real time and then runs these data through a model that predicts driving risks. The DRVR system uses facial expression-based driver attention and fatigue models to 'provide diverse in-process risk management solutions' meant to avert accidents and subsequent insurance-claim filings. A representative of Ping

An's Property and Casualty Insurance Technology Center revealed that, in addition to driver facial expression data, the cars are outfitted to incorporate real-time data on other cars and the roads being used.[113] This real-time analysis can allegedly catch drivers 'dozing off, smoking, playing with mobile phones, carelessly changing lanes, [and] speeding'. Ping An's Chief Scientist, Xiao Jing, praised this AI system's acceleration of the insurance-claim investigation process.[114]

Emotion Recognition Outside of Cars

To date, Chinese driving-safety applications of emotion recognition capabilities tend to focus on individuals inside of cars; yet there is also emerging interest in how the technology could be used at highway toll booths. An academic paper by four researchers at the Xi'An Highway Research Institute proposes an early-warning system that would use micro-expression recognition to detect drivers likely to commit highway fare evasion.[115] The authors note that, in some parts of China, highway toll booths are already outfitted with automated licence plate readers and facial recognition-equipped cameras to track the vehicles of drivers who evade tolls. In addition to their proposal that micro-expression recognition be used to detect suspects likely to commit fare evasion, they broaden the scope to include 'early detection' of drivers who may pose a physical threat to tollbooth operators.[116] Such a system would require the creation of a facial-expression database comprising people who have evaded fares or perpetrated violence against tollbooth operators in the past, which could be shared across multiple highway systems and updated with the facial expression data it would continually amass.[117] This envisioned system would connect to existing highway-monitoring systems, and could link a driver's facial recognition and facial expression data with the same individual's licence information, creating what the authors describe as a 'highway traveller credit database' (高速公路出行者信用数据库) that could be shared with the Ministry of Public Security, as well as with transportation and security-inspection authorities, as evidence of fare evasion.[118] While there has been no indication that this particular project is to be trialled or implemented thus far, there are some signs of state support for the development of emotion recognition for driving safety.

State and Tech Industry Interest

The city of Guangzhou issued a suggested research topic, 'Research and Development on Applications of Video Surveillance Inside and Outside of Vehicles', in its 2019 special application guide for 'smart connected cars'. Specifically, the application guide expressed interest in supporting 'recognition of and feedback on mental state, emotional conditions, vital signs, etc. to improve security in the driver's seat', and 'achievable emotion recognition of drivers, automated adjustment of the vehicle's interior aroma/music category/colour of ambient lighting to stabilize the driver's mental state'.[119]

Tech companies that have provided emotion recognition capabilities in other use cases have shown interest in the driving-safety subsector. Xinktech, for instance, has mentioned highway management as a next step for its 'Lingshi' multimodal emotion analysis system.[120] The company is also researching in-car emotion recognition for potential use in taxis. In making a case for studying the emotional expressions of taxi drivers and passengers before and after incidents of verbal and physical conflict erupt, Xinktech CEO Ling Zhihui cited a series of murders and rapes that drivers for ride-hailing company Didi Chuxing committed. Ling suggests these data can be used to train early-warning models that would alert Didi's customer-service representatives to intervene and prevent passenger harm.[121] Much like with technologies that purport to predict acts of terror, these 'solutions' could instead cause problems for drivers incorrectly flagged as at-risk of harming a passenger. Recording suspicious behaviour in the driver's ridesharing profile, banning the driver from the platform, or escalating the situation to involve the police are all potential negative outcomes if emotion recognition is applied in this setting.

Education

Emotion and Edtech

Educational technology (edtech) makes a mix of promises about student safety, learning progress, and effective teaching approaches in its applications, although (for reasons discussed below) it is clear there is a burgeoning market for these tools across the world, and within China in particular. Face-recognition technologies used in addition to, or embedded within, edtech products have sparked debate around student privacy and the discriminatory potential of such technologies in various national contexts.[122] For instance, as the sales of test-proctoring software have skyrocketed due to COVID-19 school shutdowns, educators are alarmed at how these platforms monitor students to detect signs of cheating.[123] Due to these systems' design, darker-complexioned students have encountered technical difficulties in which systems struggle to identify their faces. Nonbinary-identifying students, as well as student parents and neurodiverse students, have also come up against problems with how test-proctoring systems identify their faces and body movements.[124]

China's Push for 'AI+Education'

Pre-pandemic uses of face and facial expression recognition in schools tended to fall into three general purposes: conducting face-based attendance checks, detecting student attention or interest in a lecture, and flagging safety threats. There are two main reasons for the push for edtech at the local and regional levels in the Chinese education sector. First, promises of progress tracking, exam preparation, and higher quality of learning resonate strongly with parents, who are willing to spend large amounts of money in a competitive education system predicated on standardised testing.[125] Second, state- and national-level AI ambitions and education-focused incentives, in particular, incentivise new edtech development and deployment.

In 2018, the Action Plan for Artificial Intelligence Innovation in Colleges and Universities stated the intention to accelerate the innovation and applications of AI in the education sector.[126] Another driver is China's New Generation AI development plan, under which local governments carry out the central government's vision of AI innovation through subsidies and other support mechanisms that enable rapid AI adoption throughout a variety of institutions.[127] The fervour for personalised learning and other technology-enhanced education outcomes in plans such as 'China Education Modernization 2035' (中国教育现代化2035) adds to the nationwide impetus.[128] As one paper on the role of private companies in the edtech sector in China posits, 'it is this pursuit of marketable products that appears to define the general approach of the private sector, rather than any underlying educational rationale for the design and development of AI applications'.[129]

Chinese Academic Research on Emotion Recognition in Education

Emotion recognition technologies gain an additional sense of urgency in light of an accompanying trove of domestic academic research in this area. Among the dozens of research papers Chinese scholars have published about machine learning-dependent emotion recognition methods and their applications, education-related applications may be viewed as less controversial, net-positive use cases. They do not consider these technologies' potential harms to students and teachers. As demonstrated in this report, academic researchers' assumptions and arguments then reappear in marketing for commercial emotion recognition technologies – even trickling down to school administrators' own descriptions of why these technologies should be used to manage students and teachers.

Researchers have explored using emotion recognition to detect students' cognitive disabilities, and isolate specific moments that interest them, as a basis for teachers to modify their lesson

plans.[130] The presumed causal link between content taught and students' outward expressions of interest are the foundations of an argument for personalised learning that many firms (described in China's Market for Emotion Recognition in Education) repeat. Another study applies deep-learning algorithms to identify students' real-time facial expressions in massive open online courses (MOOCs). [131] The authors believe emotion recognition benefits MOOCs in particular because teachers are not co-located with their students and need to enhance the quality of student–teacher interaction.[132] Although at least one study acknowledges that equating students' facial emotions with states of learning engagement is a highly limited approach, the main response to this shortcoming has been to create new versions that learn from past data (or, 'iterate') on unimodal emotion recognition with multimodal alternatives.[133]

One multimodal study of Chinese MOOC participants collected facial-image and brainwave data to create a novel dataset, comprised of Chinese learners (as opposed to human research subjects of other ethnicities), and to address low course-completion and participation rates in MOOCs.[134] Others investigated methods for using Tencent's near-ubiquitous messaging app, WeChat, to conduct face, expression, and gesture recognition that would be implemented in classrooms as a continuous, cost-efficient alternative to end-of-term questionnaire evaluations of teachers.[135] In a similar vein, another paper suggests vocal tone-recognition technology can be used like a 'smoke alarm': if teachers' voices express insufficient warmth or affinity (亲和力), they can receive reminders to do so.[136]

Academic literature within China does not touch on an important consideration in the use of emotional recognition in schools: recent research has found that current facial-emotion recognition methods demonstrate subpar performance when applied to children's facial expressions.[137] Nonetheless, as the 11 companies in Table 2 demonstrate, physical and virtual classrooms across China are test sites for emotion recognition in education. As the COVID-19 pandemic has popularised 'blended learning' (混合学习) – where some classroom instruction is conducted using digital technologies, while the rest retains the traditional face-to-face approach – several of these companies are prepared to absorb new demand.

China's Market for Emotion Recognition in Education

Given how similar emotion recognition product offerings in the education field are, one way to differentiate between them is to examine how they came to incorporate emotion recognition into their core offerings. One set is companies that did not start out in the education sector but later developed their emotion recognition software and/or hardware for education use cases (Hanwang, Hikvision, Lenovo, Meezao, Taigusys Computing). Another is edtech firms with emotion recognition capabilities ostensibly built in-house (Haifeng Education, New Oriental, TAL, VIPKID). The third comprises partnerships between edtech firms and major tech companies specialising in emotion, face, voice, and gesture recognition (EF Children's English, VTron Group). As with security applications, user interfaces from these companies illuminate which data points are used to restructure the learning experience.

As of December 2017, Hanwang Education supplied at least seven schools around China with its CCS.[138] In an interview for *The Disconnect*, Hanwang Education's general manager logged into the CCS user account of a teacher at Chifeng No. 4 Middle School in the Inner Mongolia Autonomous Region to demonstrate the app. [139] Aside from behavioural scores, the app breaks down percentages of class time spent on each of the five behaviours it recognises, and compares individual students with the class average. For example, a student who was marked as focused 94% of the time in his English class, but was recorded as only answering one of the teacher's questions in a week, was considered to have low class participation.[140]

Table 2: Companies Providing Emotion Recognition for Education

Company	Type of Instruction	Product Name and Description
EF Children's English 英孚少儿英语	In person and online	Partners with Tencent Cloud to conduct image, emotion, and voice recognition, and receives curriculum design assistance to EF's product-development teams and teachers.[141]
Hanwang Education 汉王教育	In person	Class Care System (CCS) cameras take photos of whole classes once per second, connect to a programme that purportedly uses deep-learning algorithms to detect behaviours (including 'listening, answering questions, writing, interacting with other students, or sleeping') and issue behavioural scores to students every week. Scores are part of a weekly report that parents and teachers access via the CCS mobile app.[142]
Haifeng Education 海风教育	Online	Cape of Good Hope multimodal emotion recognition platform tracks students' eyeball movements, facial expressions, vocal tone, and dialogue to measure concentration.[143]
Hikvision 海康威视	In person	Smart Classroom Behaviour Management System integrates three cameras, positioned at the front of the classroom, and identifies seven types of emotions (fear, happiness, disgust, sadness, surprise, anger, and neutral) and six behaviours (reading, writing, listening, standing, raising hands, and laying one's head on a desk).[144] Cameras take attendance using face recognition, and scan students' faces every 30 seconds.[145]
Lenovo 联想	In person	'Smart education solutions' include speech, gesture, and facial emotion recognition.[146]
Meezao 蜜枣网	In person	Uses facial expression recognition and eye-tracking software to scan preschoolers' faces over 1,000 times per day and generate reports, which are shared with teachers and parents.[147] Reports contain data visualisations of students' concentration levels at different points in class.[148]
New Oriental 新东方	Blended learning	AI Dual Teacher Classrooms contain a 'smart eye system based on emotion recognition and students' attention levels', which the company says can also detect emotional states, including 'happy, sad, surprised, normal, and angry'.[149] A subsidiary, BlingABC, offers online teaching tools such as the AI Foreign Teacher, which contains face- and voice-recognition functions. BlingABC counts how many words students speak, along with data about students' focus levels and emotions, and claims reports containing this combination of data can help students, parents, and teachers zero in on exactly which parts of a lesson a student did not fully understand.[150]
Taigusys Computing 太古计算	In person	Collects data from three cameras, one each on students' faces, teachers, and a classroom's blackboard. The system detects seven emotions (neutral, happy, surprised, disgusted, sad, angry, scared) and seven actions (reading, writing, listening, raising hands, standing up, lying on the desk, playing with mobile phones).[151]

Tomorrow Advancing Life (TAL)	Online	Xueersi Online School provides extracurricular instruction for elementary, junior high, and high school students.[152] It and other TAL online learning platforms incorporate the Magic Mirror System, which identifies 'concentration, doubt, happiness, and closed eyes' based on facial expressions. A display monitor relates information about students' attentiveness to teachers in real time, and reports are generated for parents to track their children's learning progress.[153]
VIPKID	Online	Conducts face and expression recognition on students and teachers. Generates reports to share with teachers and parents.[154]
VTron Group 威创集团	In person	Partners with Baidu and Megvii to develop face, emotion, and voice recognition technology to monitor preschoolers and generate reports for their teachers and parents.[155]

Taigusys Computing (太古计算), which has produced emotion recognition platforms for prison surveillance and police interrogations, (see the Public Security section), has a teacher user interface that displays 'problem student warnings' with corresponding emotions, such as sadness and fear. Other data visualisations combine data on expression and behaviour recognition alongside academic performance to typologise students. For instance, the 'falsely earnest type' is assigned to a student who 'attentively listens to lectures [but has] bad grades', while a 'top student' might be one with 'unfocused listening, strong self-study abilities, but good grades'.[156]

Although most of these systems are developed solely within companies, a few draw from academic partnerships and funding of smaller startups. Some of the support for emotion, gesture, and face recognition in products from one of China's biggest edtech suppliers, TAL, comes from its Tsinghua University–TAL Intelligent Education Information Technology Research Center, and from technology TAL has acquired through FaceThink (德麟科技), an expression recognition startup it has funded.[157] When it comes to selling and implementing products, several of the companies examined here have been known to play to two narratives that surpass education: parents' fears about students' safety, and 'digital divide' concerns that less-developed regions of China will technologically lag behind wealthier coastal provinces.

Tech companies use slightly different arguments for emotion recognition depending on students' age group and whether the technology is to be used for online teaching or in-person classrooms. Companies that have produced online teaching platforms for nursery school-aged children, for example, market their products as critical to not only assessing young children's budding abilities to concentrate and learn but also protecting these students' safety. Meezao (蜜枣网), which won awards from Microsoft Research Asia for its applications of emotion recognition technology in retail and banking before turning to the education field, provides one example. [158]

Meezao's founder, Zhao Xiaomeng, cited the November 2017 RYB incident, in which Beijing kindergarten teachers were reported to have abused students with pills and needles, as having made him 'recognise [that] emotional intelligence [technology's] reflection of children's emotional changes can help administrators more accurately and quickly understand hidden dangers to children's safety, such that they can prevent malicious incidents from happening again'.[159] Zhao described a trial of Meezao's technology in a preschool classroom, where the software identified fear on many of the children's faces when a male stranger with an athletic build entered the classroom.[160]

Similarly, according to VTron Group (威创集团) CEO, Guo Dan, their collaborations with teams from both Baidu and Megvii enables the use of:

> *"AI cameras, recognition algorithms, and big data analysis, to accurately obtain information on individuals' identities and teachers' and students' emotions, and to provide complete solutions for [ensuring] kindergarteners can be picked up [from school] safely, for teachers' emotional guidance, and for early warning mechanisms [that detect] child safety crises".*[161]

The faulty assumptions that these security arguments are based on remain unchallenged in the Chinese literature. As with narratives about ameliorating education across rural and lower-resourced regions of China, the companies make promises the technology alone cannot deliver on – and, indeed, are not held accountable for upholding.

Hardware giant Lenovo has extended its emotion recognition capabilities (originally used in customer-service settings) to Chinese classrooms.[162] Lenovo has sold edtech to elementary and high schools in Sichuan, Tibet, Shandong, and Yunnan (among at least a dozen provinces the company contracts with), touting sales to these provinces as a means of closing the digital divide. However, because Lenovo's emotion recognition feature is modular, it is difficult to pinpoint exactly how many of these schools use it.[163] New Oriental (新东方), which has brought its AI Dual Teacher Classroom (AI双师课堂) to over 600 classrooms in 30 cities across China, strategically spotlights its efforts in cities like Ya'an in Sichuan province.[164] Despite these sizeable user bases, in-depth testimonials of how these technologies are viewed within schools are scarce. One exception comes from the country's best-documented – and perhaps most contentious – implementation of emotion recognition, at a high school in the coastal tech hub of Hangzhou.

Public criticism directed at various applications of emotion recognition in Chinese schools does not appear to impede the likelihood that more domestic companies will apply voice and facial expression-based emotion recognition in the education sector. Factors that contribute to this potential proliferation include the breadth of market opportunities, both within and beyond schools; perceptions of technological prestige, attributed to specific institutions and the country as a whole, for leading the adoption of these tools; and local governments' policy support and subsidies of these technologies' installation and upkeep.

Emotion Recognition in Online and In-Person Classrooms

In May 2018, Hangzhou No. 11 Middle School held a 'smart campus' seminar where it unveiled a Smart Classroom Behaviour Management System (智慧课堂行为管理系统), which the world's biggest surveillance-camera producer, Hikvision, produced in conjunction with the school.[165] Computer monitors near teachers' desks or lecture stands displayed the system's assignment of the values A, B, and C to students, based on emotional and behavioural indicators, and included a column representing 'school-wide expression data'. According to the school's administration, the only behaviour registered as 'negative' was resting one's head on a desk; and if a student did this often enough to surpass a preset threshold, they were assigned a C value. Twenty minutes into each class, teachers' display screens issued notifications about which students were inattentive. These notifications disappeared after three minutes.[166] Outside of classrooms, monitors showing how many students were making 'sour faces' (苦瓜脸) and neutral faces were mounted in hallways.[167] Some media accounts in English and Mandarin suggest the technology has since been scaled back, while others indicate it has been removed altogether. Yet, in its brief trial period, the Smart Classroom Behavioural Management System revealed how perceptions of emotion recognition changed social dynamics in schools.

Students' Experiences of Emotion Recognition Technologies

Despite avowals from companies such as Meezao and Hikvision that their emotion recognition applications were designed in conjunction with the schools that use them, students appeared to have been left out of these consultations. As a Hanwang Education technician put it: "We suggest the schools ask for the students' consent before using CCS […] If they don't, there's nothing we can do."[168] Of the few students interviewed about their experiences of emotion recognition technologies in Chinese classrooms, none supported their schools' use of these systems.

At Hangzhou No. 11, which claims to have only tested the Hikvision Smart Classroom Behaviour Management System on two tenth-grade classes, some students were afraid when their teachers demonstrated the technology.[169] While one student's fear was grounded in her understanding of how powerful Hikvision's high-resolution surveillance cameras are, others worried about being academically penalised if any of their movements were recorded as unfocused.[170] "Ever since the 'smart eyes' have been installed in the classroom, I haven't dared to be absent-minded in class," reflected one student at Hangzhou No. 11.[171] This narrative can fuel belief in the power of a technology that potentially exceeds what it is being used for; one student at Niulanshan First Secondary School in Beijing was anxious that data about the moments when he is inattentive in class could be shared with universities he wants to attend.

Examples of behaviour changes in students bear out a concern that Chinese academics have regarding emotion recognition; namely, that students will develop 'performative personalities' (表演型人格), feigning interest in class if this becomes another metric on which their academic abilities are judged.[172] Some students found staring straight ahead was the key to convincing the system they were focused.[173] Experts who agree that the cameras elicit this performative instinct, however, are not in agreement about how to respond. Shanghai Jiaotong University professor Xiong Bingqi castigates the cameras as a "very bad influence on students' development […] [that] take[s] advantage of students" and should be removed.[174] He Shanyun, an associate professor of education at Zhejiang University, believes the 'role-playing' effect could be mitigated if classroom applications of emotion recognition are not tied to rewards and disciplinary measures against students, and are only used for follow-up analysis of students' learning progress. Shanghai University of Finance and Economics law professor, Hu Ling, emphasised that schools needed to do the work to convince parents and students that the technology was not being used to assess academic performance.[175] Yet, to place the onus for seeking consent on the school alone absolves the companies of responsibility.

Niulanshan First Secondary School teamed up with Hanwang to use the company's CCS cameras to conduct facial sampling of students every few months to account for changes in their physical appearance.[176] This continual sampling – coupled with accounts from students at Hangzhou No. 11, who found their school's face-recognition-enabled Hikvision cameras often failed when they changed hairstyles or wore glasses – suggests this converse scenario of error-prone cameras both undermines the argument that these new technologies are fool proof and can even lead to students being apathetic about these new measures.[177] At Hangzhou No. 11, some students noticed attentive classmates were sometimes mislabelled 'C' for unfocused behaviour.[178] Perception of this error led these students to discredit the system, with one commenting: "it's very inaccurate, so the equipment is still continually being debugged," and another admitting: "We don't look at this thing too often."[179]

Perceptions of inaccuracies do not always end with ignoring technology, however. Some Chinese academics see the misjudgements of emotion

recognition systems as merely a function of insufficient data, therefore requiring additional data collection for improvement. For instance, Professor He posited that a silent, expressionless student's cognitive activity may not be legible to a facial emotion-reading system, supporting a 'need for further empirical research and real-world applications to explain the relationship between [facial] expression and learning results.'[180] This support for more emotion recognition experiments, rather than an interrogation of the premise that emotion recognition is appropriate in classrooms, is shared among education experts who advise the government. In a panel that Hangzhou No. 11 held to seek expert opinions on its applications of Hikvision's Smart Classroom Behaviour Management System, Ren Youqun – East China Normal University professor and Secretary-General of the Education Informatization Expert Committee of China's Ministry of Education – echoed Professor He's call for more research while the technology is still immature.[181] Headmaster of Qingdao Hongde Elementary School and edtech expert, Lü Hongjun, concurred – with the caveat that these technologies should only be rolled out experimentally in some classrooms, rather than becoming omnipresent in schools and placing too much pressure on students.[182] Finally, frustration with emotion recognition in schools has cropped up in classrooms. According to Hanwang's Zhang, students at Chifeng No. 4 Middle School unplugged the system's cameras the day before their final exams.[183]

Students, of course, are not the only ones whose classroom experiences have been reconfigured by the introduction of emotion recognition technologies. The managerial nature of these technological interventions extend to teachers, whether they are treated as tools to measure teachers' performance or as a disciplinary aid to draw teachers' attention to distracted students.

Teachers' Experiences of Emotion Recognition Technologies

Almost all the companies working on applying emotion recognition in educational settings claim the data they generate on students' attention levels and emotional states can be used to make teachers more effective at their jobs. The implicit and explicit pitches that emotion recognition vendors make about the technology's benefit to teachers echo the Chinese research literature, which equates facial reactions to course content with interest in the material. Statements about students' emotional responses inevitably become a commentary on teachers' performance.

New Oriental characterises facial expressions as representing 'students' true feedback' to teachers' instruction.[184] Comparing traditional classrooms to 'black boxes', where 'the quality of teaching could not be quantified', one account of TAL's Magic Mirror claims teachers can obtain effective suggestions to improve their instruction methods from the reports the product derives.[185] Haifeng Education depicts its Cape of Good Hope platform as capable of student-concentration monitoring, 'course quality analysis', and reduction of teachers' workloads. As in the papers studying how to apply emotion recognition to MOOCs, Haifeng suggests teachers can consult their platform's real-time data visualisation of students' emotional responses to a lecture, and judge how to adjust their teaching in response.[186] A write-up of Hangzhou No. 11's Hikvision collaboration in *ThePaper* (澎湃) likewise maintained that a teacher's popularity can be determined through comparing data on the number of 'happy' students in their class to that of another teacher who lectures the same students.[187] A representative of VIPKID, a popular online English teaching platform that hires foreign teachers to provide remote instruction, noted that, 'through facial expressions, parents can understand the state of their children's learning. We can also closely follow foreign teachers' teaching situation at any time.'[188]

At the same time, promises about advancing personalised learning and improving teacher quality fail to elaborate on what kinds of recommendations teachers are given to achieve these vague outcomes. For example, there is no clear differentiation regarding whether data on which students were 'attentive' during certain parts of a lecture reflects interest in the material, approval of a teacher's pedagogical methods, or another reason altogether – let alone guidance on how the data can be converted into personalised solutions tailored to each student.

In the aforementioned expert panel on the use of the Smart Classroom Behaviour Management System at Hangzhou No. 11, the difficulty of balancing competing interests and drawbacks to teachers and students was evident. Ni Mijing, a member of the national Committee of the Chinese People's Political Consultative Conference and deputy director of the Shanghai Municipal Education Commission, acknowledged the value of data on students' reactions to evaluations of teachers, and advocated openness to edtech trials as a way for schools to learn from their mistakes.[189] However, he also warned:

> *"We should oppose using technology to judge the pros and cons of children's study methods, [and we] should even oppose using this set of technologies to judge teachers' teaching quality, otherwise this will produce data distortion and terrible problems for education […] Excessive top-level design and investment are very likely to become a digital wasteland, a phenomenon I call a digital curse."*[190]

As with students who expressed doubts about the accuracy of Hikvision's Smart Classroom Behaviour Management System and other Hikvision face recognition cameras at their school, teachers have implied the technology has not changed much about how they do their work. For example, one teacher at Hangzhou No. 11 said: "Sometimes during class I'll glimpse at it, but I still haven't criticised students because their names appear on it."[191] Chinese news reporting has paid little attention to teachers' impressions of emotion recognition, focusing more on students, as well as on the biggest champions of these technologies: school administrators.

School Administrators' Perceptions of Emotion Recognition Technologies

School administrators have tended to take defensive stances on the value of emotion recognition technology to their schools. The defensiveness is, in part, a response to spirited public discussion about privacy concerns surrounding the use of emotion recognition in schools in China. For instance, Megvii's implementation of an attendance-, emotion-, and behaviour-recognition camera, the MegEye-C3V-920, at China Pharmaceutical University in Nanjing met with criticism on Chinese social media.[192] While social media commentary focuses on a suite of privacy and rights violations, news media accounts instead tend to focus on the risks of data leakage and third-party misuse.[193]

Hangzhou No. 11 Principal, Ni Ziyuan, responded to criticisms that the Hikvision-built Smart Classroom Behaviour Management System violates student privacy with the rebuttal that it does not store video recordings of classroom activity, but instead merely records the behavioural information extracted from video footage.[194] Vice Principal Zhang Guanchao has also tried to assuage privacy concerns by pointing out that students' data are only stored on local servers (rather than in the cloud), supposedly preventing data leaks; that the school's leadership and middle management have differentiated permissions for who can access certain student data; and that the system only analyses the behaviour of groups, not individuals.[195] Hanwang Education's CEO maintains the company's CCS does not share reports with third parties and that, when a parent is sent still images from classroom camera footage, all students' faces except their child's are blurred out.[196] In general, the defences that administrators have raised ignore the concerns that students and education experts have voiced about these technologies.

Other administration accounts of the Hikvision system at Hangzhou No. 11 stand in questionable contrast with students' and teachers' comments. In an interview for *ThePaper*, Vice Principal Zhang boasted the system had achieved positive results: students were adapting their behaviour to it, and teachers used data about students' expressions and behaviour to change their approach to teaching.[197] While Hangzhou No. 11's principal and vice principal said facial expression and behavioural data would not affect evaluations of students' academic performance, in an interview with the *Beijing News*, Zhang said: "Right now [I] can't discuss [whether this system will extend to] evaluations."[198]

Despite administrators' ardent defences of the Smart Classroom Behaviour Management System, one account suggests its use was halted the same month it was launched.[199] In spring 2019, Vice Principal Zhang announced the school had modified its system to stop assessing students' facial expressions, although the cameras would still detect students resting heads on desks and continue to issue behavioural scores.[200] The contradictory statements the administration issued, along with this retraction of facial expression detection, may point to a mismatch between expectations and reality when it comes to applying emotion recognition in schools.

Positive media coverage of schools' embrace of new technologies prevail over accounts of the ultimate underuse and distrust of emotion recognition technologies in the education sector. Moreover, school administrators continue to benefit from touting these technological acquisitions when publicising themselves to local government authorities as progressive and worthy of more funding. On a national level, being the first country to publicly trial these technologies is a source of pride. For instance, one account of TAL's Magic Mirror posited that 'emotion recognition technology is also China's "representative product of independent intellectual property rights"' – a description that reappears on Hangzhou No. 11's official WeChat account in a write-up of the Hikvision Smart Classroom Behaviour Management System.[201]

At a local level, policymakers' guidance is more directed. The May 2018 event, at which the Hangzhou No. 11–Hikvision collaboration was first launched, was organised by the Hangzhou Educational Technology Center – itself supervised by the Hangzhou Education Bureau. The Hangzhou Educational Technology Center is in charge of both edtech procurement and technical training for primary and secondary schools in the city.[202] While Hangzhou is among China's wealthier cities, with resources at its disposal to conduct edtech experiments, the user bases of the aforementioned tech companies are likely to grow, leading more of them to come up against the same issues Hangzhou No. 11 did. Not all municipal and provincial governments neglect public responses to these technological interventions; Shenzhen's Municipal Education Bureau decided against implementing online video surveillance of kindergarten classrooms to protect student privacy.[203] Examples like this are the exception, however, and do not preclude other cities and provinces from experimenting with emotion recognition.

A central tension that schools will continue to face concerns whether emotion recognition will be used to measure academic performance, student behaviour, or both. 'Function creep' –technologies' expansion into collecting data and/or executing functions they were not originally approved to collect or execute – is another possibility. For example, in acknowledging that Hangzhou No. 11's Smart Classroom Behaviour Management System may label students who rest their heads on their desks due to illness as 'inattentive', Vice Principal Zhang suggested the school nurse's office could establish 'white lists' of ill students to prevent them from being unfairly marked as unfocused in class.[204] Similarly, Hangzhou No. 11 implemented facial recognition as a form of mobile payment authentication in its cafeteria in 2017. Not long after, the school used face recognition to monitor library loans and compile annual nutrition reports for each student, which shared information about students' cafeteria food consumption with their parents.[205]

Parents' Perceptions of Emotion Recognition Technologies

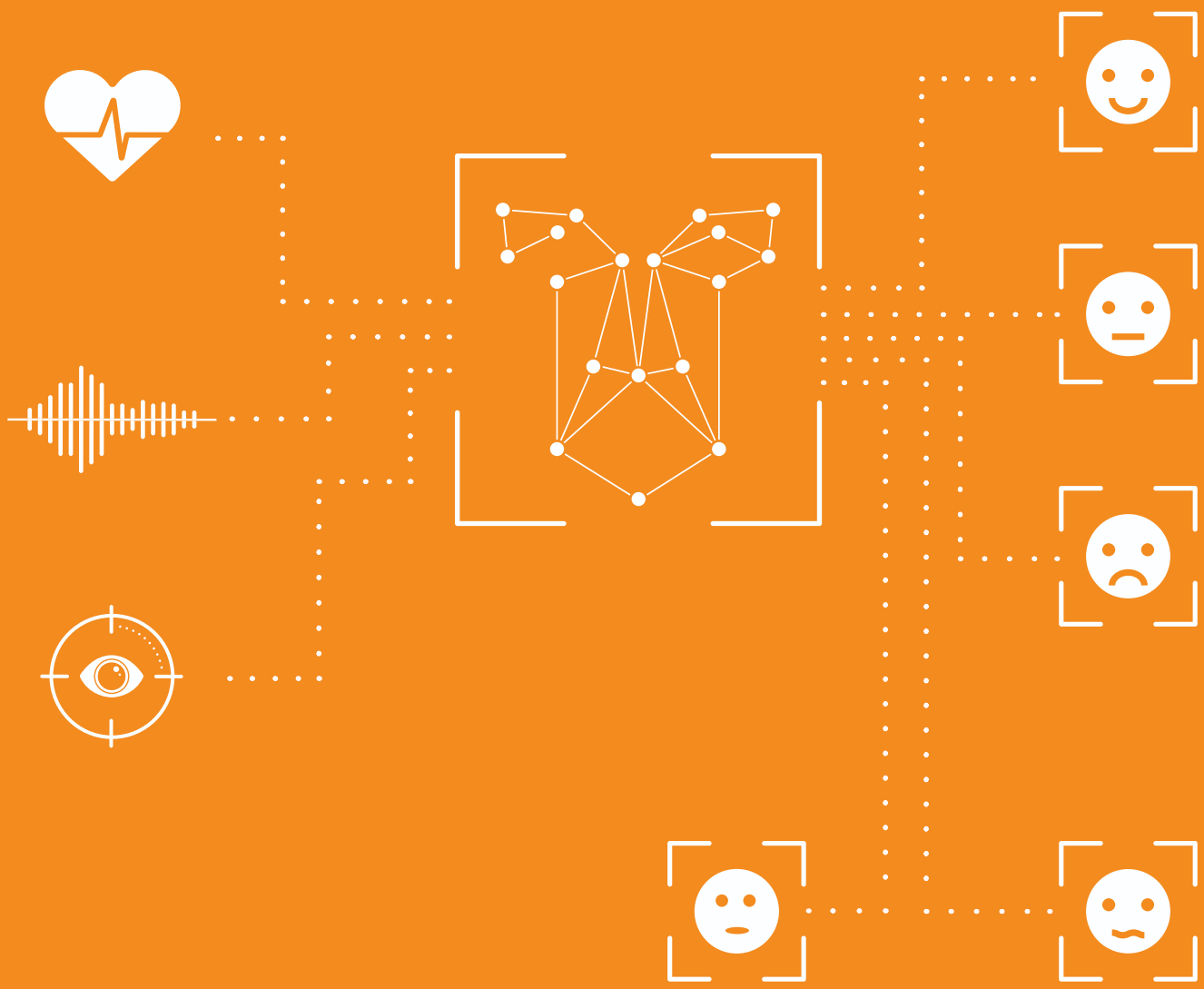
Although parents can – in theory –advocate for their children's interests in schools, the extent to which they have done so regarding schools' use of emotion recognition is unclear. One article, reporting on a Chinese Communist Party-sponsored technology expo that featured TAL's Magic Mirror, quoted an impressed parent who felt the use of this technology made their child's education better than that of their parents' generation.[206] Yet, a blog post declared that parents disliked this monitoring of their children, and that some companies subsequently removed phrases like 'emotion recognition', 'facial recognition', and 'magic mirror' from their marketing.[207]

Regardless of parents' views on the issue, Professor Hu Ling of Shanghai University of Finance and Economics noted that "schools hold the power to evaluate, punish, and expel", and so "parents won't sacrifice the students' futures by standing up against the schools, which leaves the students in the most vulnerable position."[208] Companies, too, wield power over parents. In discussing the appeal of their product, Hanwang Education's CEO commented on Chinese parents' vigilance over their children's academic performance and behaviour in class as a product of the national education system's focus on testing as a central determinant of future opportunities.[209]

Professor Hu hit upon the question that schools will continue to revisit regarding not only emotion recognition, but also all future technological interventions that purportedly make education more efficient, effective, quantifiable, and manageable:

> *"The most fundamental question is, what do we expect education to become? If it is guided by efficient test-taking, it will naturally cut all classroom behaviour into fragments, layers, and scores, [and] an algorithm will evaluate if you are a child who loves to learn or if you are a child who doesn't love to learn."*[210]

3. Emotion Recognition and Human Rights



International human rights are guaranteed by the Universal Declaration of Human Rights and given binding legal force through the International Covenant on Civil and Political Rights (ICCPR) and in regional treaties.

States are under binding legal obligations to promote, respect, protect, and guarantee human rights in these treaties. They are also under the obligation to provide guidance to businesses on how to respect human rights throughout their operations.[211]

Private companies also have responsibility to respect human rights; the *Guiding Principles on Business and Human Rights* provide a starting point for articulating the role of the private sector in protecting human rights in relation to digital technologies. Even though these principles are not binding, the UN Special Rapporteur on Freedom of Expression and Opinion has stated that 'the companies' overwhelming role in public life globally argues strongly for their adoption and implementation'.[212]

While we have discussed the dubious scientific foundations that underpin emotion recognition technologies, it is crucial to note that, emotion recognition technologies serve as a basis to restrict access to services and opportunities, as well as disproportionately impacting vulnerable individuals in society. They are therefore fundamentally inconsistent with international human rights standards, described in this chapter.

Human dignity underpins and pervades these human rights instruments.[213] As stated in the Preamble to the ICCPR, '[human] rights derive from the inherent dignity of the human person', which is underscored by the fact that it [dignity] is not a concept confined to preambulatory clauses alone but is also used in context of substantive rights.[214] Emotion recognition strikes at the heart of this concept by contemplating analysing and classifying human beings into arbitrary categories that touch on the most personal aspects of their being. Overarchingly, the very use of emotion recognition imperils human dignity and, in turn, human rights – particularly given the discriminatory and discredited scientific foundations on which this technology is built.

Right to Privacy

Emotion recognition technologies require collecting sensitive personal information for both training and application. Individuals being identified, analysed, and classified may have no knowledge that they are being subject to these processes, making the risks that emotion recognition poses to individual rights and freedoms grave. While these technologies in isolation do not necessarily identify individuals, they can be used to corroborate identities when used among other technologies that carry out identification. This significantly impedes ability to remain anonymous, a key concept in the protection of the right to privacy as well as freedom of expression.[215]

A common thread across all use cases discussed in this report is the concentration of state and industry power; to use emotion recognition technologies, companies and state actors have to engage in constant, intrusive, and arbitrary qualitative judgements to assess individuals. It is important, therefore, to consider surveillance as an inevitable outcome of all emotion recognition applications. For example, all the use cases for early warning, closer monitoring, and interrogation related to public security are deployed on the grounds that they are necessary to prevent crime and ensure safety. In practice, however, they are deployed indiscriminately for fishing expeditions that are unrelated to the needs of a particular operation. Mass surveillance thus increasingly becomes an end in and of itself. Further, the stated purpose of driving-safety applications is to ensure driver and passenger safety, but the outcome includes systematic invasions of privacy and significant mission creep, in the case of biometric information potentially being used for insurance purposes. A basic tenet of international human rights law is that rights may not be violated in ways that confer unfettered discretion to entities in power, which is a feature of – not a bug in – these technologies.

Any interference with the right to privacy must be provided by law, in pursuit of a legitimate aim, and necessary and proportionate.[216] Privacy concerns over biometric mass surveillance have received dedicated attention in the last few years. In a 2018 report on the right to privacy in the digital age, the UN High Commissioner for Human Rights, while discussing significant human rights concerns raised by biometric technologies, stated:

> *"Such data is particularly sensitive, as it is by definition inseparably linked to a particular person and that person's life, and has the potential to be gravely abused […] Moreover, biometric data may be used for different purposes from those for which it was collected, including the unlawful tracking and monitoring of individuals. Given those risks, particular attention should be paid to questions of necessity and proportionality in the collection of biometric data. Against that background, it is worrisome that some States are embarking on vast biometric data-based projects without having adequate legal and procedural safeguards in place."*[217]

As noted by the UN Special Rapporteur on Privacy, 'evidence has not yet been made available that would persuade the [Special Rapporteur] of the proportionality or necessity of laws regulating surveillance which permit bulk acquisition of all kinds of data including metadata as well as content'.[218]

Importantly, the nature of these technologies is also at odds with the notion of preserving human dignity, and constitutes a wholly unnecessary method of achieving the purported aims of national security, public order, and so on (as the case may be). While international human rights standards carve out national security and public order as legitimate justifications for the restriction of human rights, including privacy, these situations do not give states free rein to arbitrarily procure and use technologies that have an impact on human rights; nor do they permit states to violate rights without providing narrowly tailored justifications and valid, specific reasons for doing so.[219]

Right to Freedom of Expression

Freedom of expression and privacy are mutually reinforcing rights. Privacy is a prerequisite to the meaningful exercise of freedom of expression, particularly given its role in preventing state and corporate surveillance that stifles free expression. While freedom of expression is fundamental to diverse cultural expression, creativity, and innovation, as well as the development of one's personality through self-expression, the right to privacy is essential to ensuring individuals' autonomy, facilitating the development of their sense of self, and enabling them to forge relationships with others.[220]

Claims that emotion recognition technology can infer people's 'true' inner states, and making decisions based on these inferences has two significant implications for freedom of expression. First, it gives way to significant chilling effects on the right to freedom of expression – the notion of being not only seen and identified, but also judged and classified, functions as an intimidation mechanism to make individuals conform to 'good' forms of self-expression lest they be classified as 'suspicious', 'risky', 'sleepy', or 'inattentive' (depending on the use case). Second, given the wide range of current applications, it normalises mass surveillance as part of an individual's daily life, in public and private spaces. Proposed uses, such as the research paper suggesting deployment of emotion recognition technology to identify people entering Tibet who have pro-Tibetan independence views, create a dangerously low threshold for authorities to misidentify self-incriminating behaviour in a region that is already over-surveilled. Importantly, freedom of expression includes the right *not* to speak or express oneself.[221]

Right to information is an important part of freedom of expression. This includes transparency of how state institutions are operating and making public affairs open to public scrutiny so as to enable citizens to understand the actions of their governments. The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression emphasised that:

> *"Core requirements for democratic governance, such as transparency, the accountability of public authorities or the promotion of participatory decision-making processes, are practically unattainable without adequate access to information. Combating and responding to corruption, for example, require the adoption of procedures and regulations that allow members of the public to obtain information on the organization, functioning and decision-making processes of its public administration."*[222]

Companies are also subject to transparency obligations under the *Guiding Principles on Business and Human Rights*, which require business enterprises to have in place 'Processes to enable the remediation of any adverse human rights impacts they cause or to which they contribute'.[223]

Right to Protest

The use of emotion recognition can have significant implications for right to protest, that also includes freedom of assembly, including potential discriminatory and disproportionate impacts, when deployed for the purpose of public safety and security in the context of assemblies.[224] A number of examples from around the world demonstrate the tendency of states engaging in unlawful, arbitrary, or unnecessary surveillance to identify individuals exercising their right to protest.[225] Emotion recognition adds a layer of complication and arbitrariness to an already worrying trend, given the lack of a legal basis, the absence of safeguards, and the extremely intrusive nature of these technologies. The UN Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association has stated:

> *"The use of surveillance techniques for the indiscriminate and untargeted surveillance of those exercising their right to peaceful assembly and association, in both physical and digital spaces, should be prohibited. Surveillance against individuals exercising their rights of peaceful assembly and association can only be conducted on a targeted basis, where there is a reasonable suspicion that they are engaging in or planning to engage in serious criminal offences, and under the very strictest rules, operating on principles of necessity and proportionality and providing for close judicial supervision".*[226]

Right Against Self-Incrimination

In public and national security use cases, emotion recognition often paves the way for people to be labelled as 'suspicious' or meriting closer inspection, and is also used at the stage of interrogation. The attribution of emotions like guilt, anger, frustration, and so on is conducted and determined by the entity deploying this technology, which collects, processes, and categorises information to make inferences that can have a detrimental impact on human rights. This runs counter to the right against self-incrimination contemplated in international human rights law. Article 14(3)(g) of the ICCPR lays down that the minimum guarantee in the determination of any criminal charge is that every person is entitled 'not to be compelled to testify against himself or to confess guilt'. This includes the right to silence.[227] Emotion recognition flips the script on this right, and is used to detect and signal guilt. Importantly, this right against self-incrimination applies to *all* stages of criminal proceedings, from the time a person is suspected to the time of conviction (if it is so proved).

Non-Discrimination

Since emotion recognition is intrinsically predicated on mass surveillance, it can have a disproportionate impact on historically disadvantaged groups. For instance, the security applications entail mass surveillance in public spaces, and lend themselves to flagging individuals who belong to historically marginalised groups, like ethnic minorities, or who find immigration and security lines more 'stressful' than others. Hence, the use of emotion recognition will lead to new fault lines along which individuals are classified, with no obligations for these fault lines to have any correlation to objective or verifiable truths. This technology is poised to lead to discrimination, as individuals who do not conform to the norms guiding discredited scientific foundations

(e.g. people of colour, transgender, and non-binary individuals) will be disproportionately surveilled, tracked, and judged.

The UN Human Rights Council has emphasised that 'automatic processing of personal data for individual profiling may lead to discrimination or decisions that otherwise have the potential to affect the enjoyment of human rights, including economic, social and cultural rights'.[228] Although profiling can lead to discriminatory outcomes in disproportionate ways regardless of the specific technology in question, this risk is even more pronounced in the case of emotion recognition, as the criteria for classification are primed for discrimination. Consequential decisions in the contexts of hiring, national security, driving safety, education, and criminal investigations are often built on the foundations of such profiling.

Other Technical and Policy Considerations

A number of additional strategic and substantive threads of analysis in the Chinese context are worth noting. We outline these thematically below to aid in effective civil society advocacy going forward.

Function Creep

The intended goal for the use of emotion recognition systems has varied between use cases, but indications of function creep beyond use cases discussed in this report already exist. Ping An Group's demonstration, from late 2019, indicates the firm's intention to move past using emotion recognition to monitor safety and avert accidents, and towards feeding into insurance assessments.[229] Meezao has already pivoted from only providing emotion recognition in schools to also offering these technologies at the China Museum of Science and Technology to collect data on children's responses to physics and chemistry experiments.[230] This function creep has happened before: in 2017, Hangzhou No. 11 introduced facial-recognition authentication for cafeteria payments, and subsequently expanded its use to monitoring library loans and nutrition reports for each student, outlining food consumption information for parents.[231]

This function creep also stems from a more general 'tech-solutionist' tendency to using new technologies to solve administrative and social problems.

Growing Chorus of Technical Concerns

There is growing recognition of the limitations of emotion recognition technologies from the developers, implementers, and individuals subject to them. Experts who advocate using emotion recognition for security, in particular, acknowledge some drawbacks to this technology. However, most of their critiques address the technical concerns of surveillers at the expense of the real-life impacts on those being surveilled. For example, Wenzhou customs officials published a research paper on automated identification of micro-expressions in customs inspections, which admits that camera-footage quality, lighting, and the added anxiety and fatigue of travel can affect how micro-expressions are produced, recorded, and interpreted.[232]

False positives are another commonly recognised issue; however, the Chinese research and security literature often attributes these to the person under surveillance deliberately feigning emotions, rather than to the system's own flaws. The most well-known of these is the 'Othello error', in which someone telling the truth unintentionally produces micro-expressions associated with liars. This is a particularly important finding, from a human rights perspective, as the overarching issues surrounding dignity, privacy, and freedom of expression seem to be precluded from public deliberation and critique of emotion recognition technologies.

Misaligned Stakeholder Incentives

Cooperation between academic research institutions, tech companies, and local state actors reveals the perceived benefits to each group of participating in the diffusion of these technologies, which is at odds with the human rights concerns arising from them. As one study of facial recognition firms in China found, companies that received training data from the government were more likely to spin off additional government and commercial software.[233] As such – and aside from procurement contracts to furnish technology for Sharp Eyes, Fengqiao, and related pre-existing government surveillance projects – emotion recognition firms may see longer-term financial opportunities and profits from these multi-institutional collaborations.

Regional and Global Impact

Throughout the literature on emotion recognition technology in China, few companies have expressed the intention of exporting their products at this phase of their development. Media coverage of EmoKit – the company that partnered with the city of Qujing, Yunnan, to pilot test its emotion recognition interrogation platform – suggested Yunnan's geographical proximity to South and Southeast Asia could be advantageous for exports to countries that comprise the OBOR and Maritime Silk Road regions.[234] While OBOR represents a terrestrial route connecting China to Europe via Central Asia, the Maritime Silk Road is the Indian Ocean-traversing counterpart that connects ports in China, South and Southeast Asia, the Middle East, and Eastern Africa. Alpha Hawkeye has allegedly supplied OBOR countries with its technology for counterterrorism and garnered interest from Southeast Asian security departments in the Philippines, Malaysia, Thailand, Myanmar, and Indonesia.[235] Publicly available data have not provided additional evidence of this, however, and the company's own media presence has dwindled in the last two years.

Yet, if the 'biometrics 3.0' framing of emotion recognition as a next step from face recognition persists – and if these firms demonstrate that emotion recognition capabilities are easily applied where face recognition cameras are already in use – the other places to watch for potential export are markets where Chinese tech companies have already sold face recognition cameras. For instance, Hikvision has provided surveillance equipment to schools in Canada, Denmark, Dubai, India, Japan, Malaysia, Pakistan, and South Africa,[236] while Huawei has provided numerous cities around the globe – including in Asia, Africa, and Latin America – with policing platforms.[237] In 2017, Huawei issued a call for proposals that included 'dialog emotion detection based on context information', 'emotional state analysis based on speech audio signal', and multimodal emotion recognition.[238]

Ethnicity and Emotion

Racial, gender-based, and intersectional forms of discrimination in biometric technologies like face recognition have been demonstrated in a wide range of academic and civil society research in the last few years. The UN Special Rapporteur on Contemporary Forms of Racism calls for 'racial equality and non-discrimination principles to bear on the structural and institutional impacts of emerging digital technologies'.[239] Criticisms of facial recognition technologies' inaccuracies across skin tone and gender map onto debates around emotion recognition, along with an additional variable: cultural differences in expressions of emotion.

With some exceptions, Chinese companies tend to tout the narrative that facial emotion expressions are universal,[240] but years of scientific evidence demonstrate cultural differences in facial expressions and the emotions they are interpreted to signify. This marketing strategy is unsurprising, however, given its ability to boost faith in the technology's alleged objectivity and to unearth 'true' emotions, while also paving a future path to its international export. Wang Liying, a technical director at Alpha Hawkeye, proclaimed that 'the entire recognition process is uninfluenced by expression, race, age, and shielding of the face'.[241]

Research suggests otherwise. In her paper 'Racial Influence on Automated Perceptions of Emotions,' Professor Lauren Rhue compiled a dataset of headshots of white and Black male National

Basketball Association (NBA) players to compare the emotional analysis components of Chinese face recognition company Megvii's Face++ software to Microsoft's Face API (application programming interface). In particular, she found 'Face++ rates black faces as twice as angry as white faces,' while Face API views Black faces as three times as angry as white ones.[242]

China is not the only country whose tech firms factor race into facial recognition and related technologies. However, its tech sector's growing influence over international technical standards-setting for these technologies presents an opportunity to address the domestically long-ignored consequences of technological racial and ethnic profiling. Instead of this open reckoning, admission of racial inequities in training datasets tends to become a justification for the creation of datasets of 'Chinese faces' to reduce inaccuracies in domestic applications of emotion recognition.[243]Arguments like this account for the potential bias of datasets that may over-represent a tacitly implied Han Chinese range of facial features and expressions while failing to address if and how new datasets created within China will draw samples from China's 56 officially recognised ethnic groups.

Some companies' open-source APIs include race variables that raise a host of concerns about human rights implications particularly for ethnic minorities – even before considering sources of training data, accuracy rates, and model interpretability. Baidu's face-detection API documentation includes parameters for emotion detection as well as race, with a sample of an API call return including 'yellow' as a type of race. Taigusys Computing's open-source expression-recognition API includes 'yellow', 'white', 'black', and 'Arabs' (黄种人, 白种人, 黑种人, 阿拉伯人) as its four racial categories. Neither company accounts for why race would be assessed alongside emotion in their APIs. This is untenable for two reasons. First, fundamental issues surrounding the discredited scientific foundations and racist legacy of emotion recognition makes the existence of such systems (and categories) deeply problematic. Second, the solution to the discriminatory effects of these systems is not to add more nuanced alternatives for categorising race, but rather to ban the use of such technologies altogether.[244]

Companies' Claims About Mental Health and Neurological Conditions

Proposed uses of emotion recognition to help people with neurological conditions, disabilities, and mental health afflictions are not new to the field. Affectiva has stated it began its work by developing a 'Google Glass-like device that helped individuals on the autism spectrum read the social and emotional cues of other people they are interacting with'.[245] While this report excludes an in-depth analysis of similar use cases, which are often carried out in medical institutions, it must take into account a critical omission in the emerging literature on commercial applications of emotion recognition in China: thus far, companies have ignored questions of how these technologies will work for neurodiverse individuals. Companies engaged in non-medical applications make particularly troubling claims about their ability to detect mental health disorders and neurological conditions (both diseases and disorders) – highly discrete categories that this literature often groups together, as though they were indistinguishable.

Chinese companies like Taigusys Computing and EmoKit have mentioned autism, schizophrenia, and depression as conditions they can diagnose and monitor using micro-expression recognition.[246] Meezao CEO Zhao said the company is testing its emotion recognition technology on children with disabilities; for instance, to detect types of smiling that could serve as early indicators of epilepsy.[247] One concern is that these systems will impose norms about neurotypical behaviour on people who do not display it in a way the technology is designed to detect.[248] Another possible issue involves potential discrimination against people the technology perceives as exhibiting such conditions.

Although Meezao's public-facing marketing materials have sidestepped the question of what emotion recognition reveals about students' mental health, a 2018 academic paper featuring the company's CEO as a co-author, entitled 'Application of data mining for young children education using emotion information', briefly touches on this topic.[249] The paper cites research that has found suicide to be the top cause of death among Chinese adolescents, and it partially attributes this to Chinese families lacking contact with their children and paying insufficient attention to their children's emotions.[250] In support of the paper's proposed emotion recognition intervention, the co-authors maintain that:

> *"Our system has the potential to help analyze incidents such as child abuse and school bullying. Since our intelligent system can help catch and analyze abnormal situation [sic] for discovering and solving problems in time, it will be easier to protect children from hurt. For instance, if a child shows persistent or extremely negative emotions, it is rational for us to pay attention to what he/she has suffered."*[251]

Of the companies that insist they can detect these conditions, none have offered explanations of how their technologies analyse emotions while taking this assessment into account; for example, how might a student with attention deficit hyperactivity disorder (ADHD) be monitored for attentiveness, compared with her classmates who do not have this diagnosis? In general, Chinese researchers and tech firms appear not to have deliberated about how differently abled and neurodiverse people will interact with emotion recognition systems built into any of the use cases explored in this report.

Emotion and Culpability

The inherent ethnic and ableist biases that may seep into emotion recognition's use can be amplified when early-warning systems flag individuals or groups who exhibit 'suspicious' emotions as deserving of additional monitoring. Although a 2016 research paper by two engineers from Shanghai Jiaotong University was met with major international criticism for perpetuating racism – the authors developed an algorithmic way to detect facial features that allegedly correlate to criminality – opposition to this type of work has, unfortunately, not hindered related uses of emotion recognition.[252] Claims that emotion recognition technologies' use in surveillance and interrogation reduce prejudice – because they are seen as faster, imperceptible to people under surveillance, and 'objective' compared with human, error-prone alternatives – detract from the greater likelihood that they will instead be more discriminatory for everyone.

Some companies unwittingly expose this reality in their appeals to law enforcement officials. A 'market pain point' that EmoKit singles out is that amendments to Chinese criminal law have *raised* standards for evidence, investigations, and trials. The company claims that, faced with these 'implementations of principles of respect and protections of human rights […] basic-level police forces are seriously inadequate, and need to achieve breakthroughs in cases within a limited period of time. [They] urgently need new technologies.'[253] The implication here is that emotion recognition technologies, like those EmoKit provides for police interrogation, can circumvent some of the new safeguards set in place to protect suspects' rights.

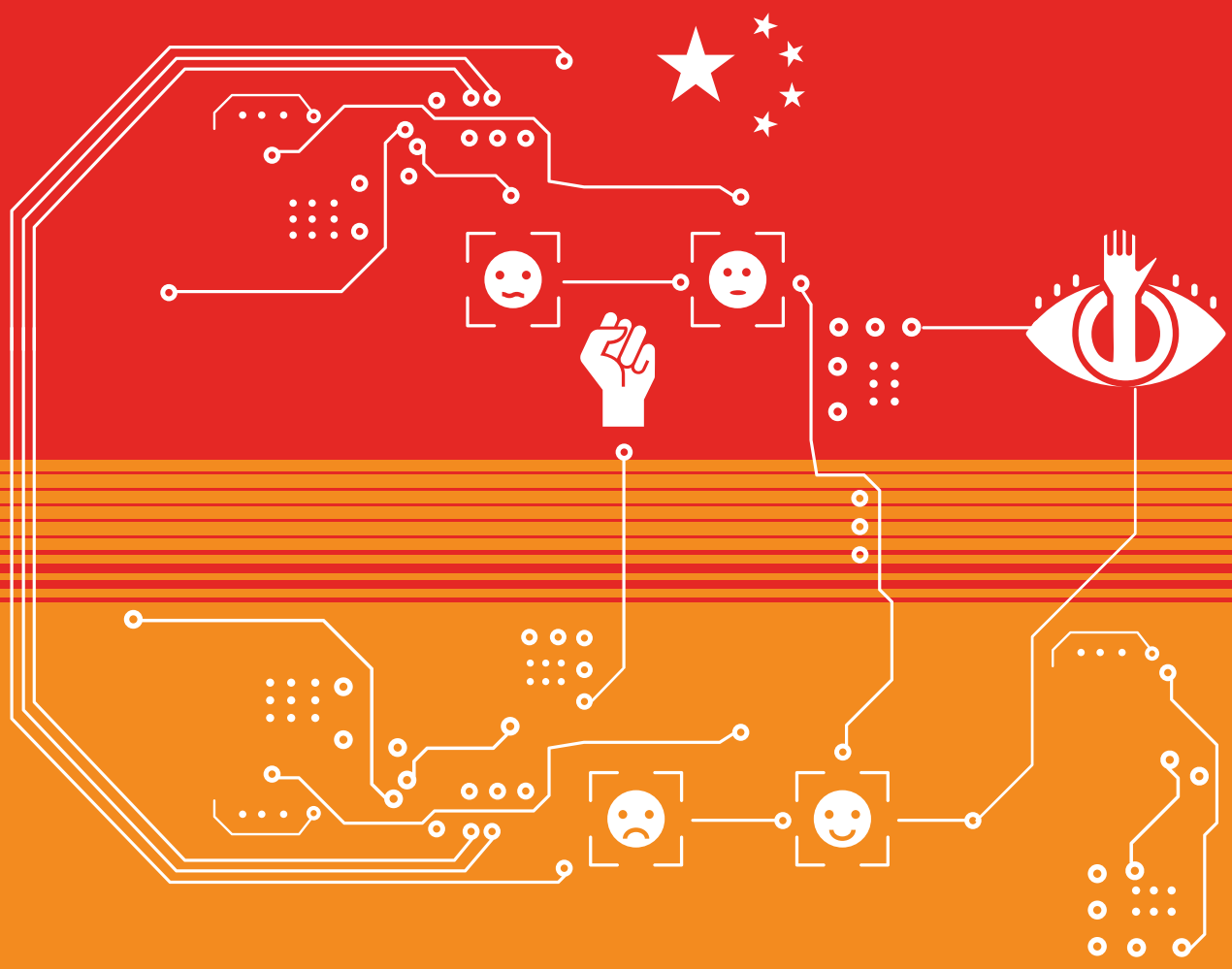
Although not all police officers view this approach as a way to get around the law, an equally problematic possibility is that some will believe that using emotion recognition in interrogation is more scientific and rights-protective. As far back as 2014, an academic paper from the People's Public Security University of China, detailing how law enforcement officials could be trained to visually observe micro-expressions, made a similar argument:

> *"Under the new Criminal Procedure Law's principle that 'no individuals can be forced to prove their guilt,' it has become more difficult to obtain confessions. Criminal suspects often respond to questioning with silence and unresponsiveness. In actuality, it is common for investigations to turn up no clues. Micro-expression analysis techniques can now solve this issue."*[254]

Specifically, investigators would be trained to introduce 'stimuli' – such as the names of people and objects related to a crime – while watching for micro-expressions that correspond to these words. They would then treat terms that elicit these minute responses as 'clues' in a case. The paper presaged the ability to return to archival interrogation video footage to search for moments when incriminating micro-expressions appeared. When AI is brought into the procedure, even more of these moments can presumably be identified. An article about Shenzhen Anshibao confirmed the technology could be used for post-mortem emotion recognition, citing video footage of the Boston marathon bombing as an example.[255]

The role of security blacklists and criminal backgrounds is also critical to the justifications that companies, researchers, and the state present for emotion recognition. Advocates of emotion recognition for public security note that, while face recognition enables cross-checking with police blacklist databases, they fail to account for people who lack criminal records. One paper, from the Public Security College of Gansu University of Political Science and Law, laments that current facial recognition systems in China lack data on residents of Hong Kong, Taiwan, Macao, and other foreign nationals. Micro-expression recognition, the authors argue, would widen the net of which 'dangerous' people can be flagged in early-warning systems.[256] This suggestion takes on added portent in light of China's recent crackdowns on Hong Kong protests and the instatement of China's new national security law there.

4. China's Legal Framework and Human Rights



China's legal landscape around data protection and AI is multi-layered and constantly evolving. Two of the main contributions of this report are:

1. Unpacking one national context – including incentives, actors, and narratives – within which these systems are meant to function; and

2. Demonstrating the fundamental incompatibility of emotion recognition systems with international human rights law.

This section lays down relevant Chinese legal standards and norms that will feed into the regulation and development of the emotion recognition market, with the aim of providing a sense of the landscape and open questions to ask in future work.

China's National Legal Framework

Relationship to International Legal Frameworks

In October 1998, China signed, but did not ratify, the ICCPR. This has been the focus of much international and national scrutiny, with several pushes for ratification; at the time of writing this report, however, it has still not been ratified.[257] Even so, China remains bound by the provisions of the ICCPR to some degree. In the 2016–2020 National Human Rights Action Plan for China, the Information Office of the State Council states:

> *"China shall continue to advance related legal preparations and pave the way for ratification of the International Covenant on Civil and Political Rights.*
>
> *China shall fully participate in the work of the UN's human rights mechanisms, and promote the United Nations Human Rights Council (HRC) and other mechanisms to attach equal importance to economic, social and cultural rights as well as civil and political rights, and function in a fair, objective and non-selective manner".*[258]

The legal preparations to ratify the ICCPR have been in motion for at least a decade, with little tangible progress.[259] It is not clear what incremental advances towards this goal are implied in the 2016–2020 National Human Rights Action Plan.

National Law

Chinese Constitution

Article 40 of the Chinese Constitution enshrines the privacy of correspondence, although this does not extend to individual data or information.[260] While Article 35 states: 'Citizens of the People's Republic of China enjoy freedom of speech, of the press, of assembly, of association, of procession and of demonstration', there is little elaboration on what this encompasses or how it is legally construed. Given that the constitution does not qualify as a valid legal basis of judicial decision or interpretation in China, its scope is decidedly limited.[261]

Even so, the pushback against unfettered collection of biometric data and mass surveillance of individuals is steadily growing through a constitutional focus. In a compelling case against the use of facial recognition in subways, for instance, a professor at Tsinghua University argued that the authorities in question did not prove the legitimacy of collecting sensitive personal information for this purpose, and invoked constitutional principles of equality, liberty, and personal freedoms.[262]

Data Protection

China's data-protection landscape is chiefly motivated by the pursuit of corporate accountability, as opposed to the protection of individual autonomy, of human rights, or against overreach of government power.[263] This stems from a generally low level of trust within the economy and increasing suspicion of fraud and cheating. The construction of safeguards and guidelines, despite drawing strong influences from the General Data Protection Regulation (GDPR), are therefore similar in form but not in incentives.

Instruments

The Chinese data-protection landscape consists of multiple instruments. At the time of writing, the interplay between these instruments is unclear, as are precise legal definitions and practical enforcement of proposed standards. The room for interpretation and executive decisions around definitions is large, which is an important consideration in dissecting this area of law.

The 2017 Cybersecurity Law is the most authoritative piece of data-protection legislation in China thus far, entering the public eye amid aggressive plans for AI development and leadership.[264] It was enacted to 'ensure cybersecurity; safeguard cyberspace sovereignty and national security, and social and public interests; protect the lawful rights and interests of citizens, legal persons, and other organizations; and promote the healthy development of the informatization of the economy and society'.[265] Within China's governance framework, and in tandem with the National Security Law and the Counterterrorism Law, the Cybersecurity Law reinforces the amalgamation of data security and national security that is pervasive throughout China's data-protection regime.[266] In general, the approach to data protection in China is risk-based, and does not stem from wider rights-based considerations. The Consumer Rights Protection Law, for instance, explicitly lays down provisions for the protection of consumers' personal information, and was the instrument through which consumers challenged Alibaba over a breach of personal data in 2018.[267]

The Draft Data Security Law, released in 2020, fleshes out regulation and specifications for the governance of data related to national security and the public interest.[268] Alongside the Data Security Law, the draft Personal Information Protection Law, released in October 2020, focuses on the protection of personal-information rights and specifically addresses the installation of image-collection and -recognition equipment, stating that such collection can only be used 'for the purpose of safeguarding public security; it may not be published or provided to other persons, except where individuals' specific consent is contained or laws or administrative regulations provide otherwise'.[269]

The Cybersecurity Law makes China's priorities around governing information and communications technologies (ICT) explicit, and gives rise to wide-ranging institutional actors, substantive areas of focus, regulations, and standards. The Cybersecurity Law has been described as sitting astride six distinct systems, which make up the evolving framework governing ICT in China.[270] The Personal Information and Important Data Protection System is one example.

The first significant set of rules for the protection of personal information, the Personal Information Security Specification, became operational in May 2018 and was revised in 2020.[271] Issued by China's national standards body, TC260, it contains guidelines for collection, storage, use, sharing, transfer, and public disclosure of personal information. 'Standards', in the Chinese context, are understood as not only technical specifications but also policy guidelines or legislation laying down benchmarks against which companies can be audited. Standards are also powerful indicators of what authorities and legislators should aspire to, both at the time of enforcement and while formulating laws. This makes them a significant piece of the data-protection puzzle in China, given the wide ambit for authorities' discretion in interpretation and enforcement.

The Specification is chiefly meant to address security challenges. According to drafters of the standard, it was modelled after the GDPR, but with important differences regarding the definition of consent (by allowing implied or silent consent) and personal-data processing in lieu of consent – a salient departure for the purposes of this report, as the intention was to be more 'business-friendly' than the GDPR and to enable the proliferation of China's AI economy, which depends on access to large datasets.[272]

Biometric Data

Biometric information is explicitly included in the definition of personal information under both the Cybersecurity Law and the Specification. However, the Specification includes it under the definition of both personal information and sensitive personal information, calling for encryption at the time of transferring and storing the latter class of data. Sensitive personal information includes the personal data of children aged 14 and younger. This creates confusion as to legitimate grounds for the collection and use of biometric data, especially given the different standards of consent required for each: while personal data requires authorised consent, sensitive personal data mandates explicit consent. Crucially, under the Specification, consent need not be obtained in cases related to national security, public safety, significant public interest, and criminal investigations, among others – all grounds that will be invoked in the use cases discussed in this report.

However, the regulation of biometric data within this evolving regime potentially goes beyond the confines of personal data. The Cybersecurity Law contemplates two broad types of data: personal information and 'important data'. Requirements for the storage, processing, sharing, and consent of data depend on how they are classified. Although 'important data' has yet to be defined clearly, one essay by the lead drafter of the Specification, Dr Hong Yanqing, states that personal data refers to autonomy and control over one's data, whereas important data affects national security, the national economy, and people's livelihoods.[273] Although a more precise definition is crucial for in-depth analysis, at first glance, biometrics falls under both categories.

The Draft Data Security Law, for instance, establishes a system for classification of data which would invoke distinct grades of data protection, contingent on the level of risk and potential severity of harm that may arise from the abuse of data in the context of, inter alia, national security, public interests, falsifying data, and so on. It also anticipates that governments and concerned agencies will define what constitutes 'important data'.

Standardisation

A number of data- and AI-related standards have cropped up in China over the last few years and, given their function as both regulatory tools and technical specifications, deserve special attention. In addition to the international standardisation efforts already discussed in this paper (e.g. ITU and the unique regulatory effect of standards like the Personal Information Security Specification), a number of developments significantly impact biometric technologies.

In 2018, the Standards Administration of China released a White Paper on AI Standardization, which recognises that, 'because the development of AI is occurring as more and more personal data are being recorded and analyzed […] in the midst of this process, protecting personal privacy is an important condition for increasing social trust'.[274] Trust seems to be a prevalent theme throughout standardisation efforts: in 2019, the China Electronics Standardization Institute released a Biometric Recognition White Paper noting the importance of standardisation in ensuring product quality and testing capabilities.[275] The State Council's New Generation Artificial Intelligence Development Plan calls for establishing an AI standards system and places immediate emphasis on the principles of security, availability, interoperability, and traceability.[276]

In line with the risk-based approach to biometric governance, TC260 released the Information Security Technology Security Impact Assessment Guide of Personal Information, which was intended to establish privacy impact assessments that lend structure to identifying risks to personal information, and which addresses both private and government actors. In forging principles for assessing impact on data subjects' rights and interests, it classifies discrimination, reputational damage, fraud, and health deterioration as high-impact risks. Discrimination, in particular, is also classified as a serious impact insofar as data subjects suffer major, irrevocable, and insurmountable impacts.[277]

Ethical Frameworks

One of the most prominent AI ethics statements to come out of China is from the Artificial Intelligence Industry Alliance, which, in 2019, published a self-discipline 'joint pledge' underscoring the need to:

> *"Establish a correct view of artificial intelligence development; clarify the basic principles and operational guides for the development and use of artificial intelligence; help to build an inclusive and shared, fair and orderly development environment; and form a sustainable development model that is safe/secure, trustworthy, rational, and responsible."*[278]

In line with priorities across regulatory tools, antidiscrimination is a prominent standard on which AI testing and development is predicated. The joint pledge calls for AI to avoid bias or discrimination against specific groups or individuals, and for companies to: 'Continually test and validate algorithms, so that they do not discriminate against users based on race, gender, nationality, age, religious beliefs, etc.' In June 2019, an expert committee set up by China's Ministry of Science and Technology put forward eight principles for AI governance, which – in line with similar efforts – underlined the importance of eliminating discrimination.[279] The committee's recommendations came on the heels of the Beijing Academy of Artificial Intelligence's Beijing AI Principles, which called for making AI systems 'as fair as possible, reducing possible discrimination'.[280]

5. Recommendations

This report has covered vast terrain: from the legacy and efficacy of emotion recognition systems to an analysis of the Chinese market for these technologies. We direct our recommendations as follows.

To the Chinese Government:

1. **Ban the development, sale, transfer, and use of emotion recognition technologies**. These technologies are based on discriminatory methods that researchers within the fields of affective computing and psychology contest.

2. **Ensure that individuals already impacted by emotion recognition technologies have access to effective remedies for violation of their rights** through judicial, administrative, legislative or other appropriate means. This should include measures to reduce legal, practical and other relevant barriers that could lead to a denial of access to remedies.

To the International Community:

1. **Ban the conception, design, development, deployment, sale, import and export of emotion recognition technologies,** in recognition of their fundamental inconsistency with international human rights standards.

To the Private Companies Investigated in this Report:

1. **Halt the design, development, and deployment of emotion recognition technologies,** as they hold massive potential to negatively affect people's lives and livelihoods, and are fundamentally and intrinsically incompatible with international human rights standards.

2. **Provide disclosure to individuals impacted by these technologies and ensure that effective, accessible and equitable grievance mechanisms** are available to them for violation of their rights as result of being targeted emotion recognition.

To Civil Society and Academia:

1. **Advocate for the ban on the design, development, testing, sale, use, import, and export of emotion recognition technology.**

2. **Support further research in this field,** and urgently work to build resistance by emphasising human rights violations linked to uses of emotion recognition.

Endnotes

1 J. McDowell, 'Something You Are: Biometrics vs. Privacy', *GIAC Security Essentials Project*, version 1.4b, 2002, https://www.giac.org/paper/gsec/2197/are-biometrics-privacy/103735. Also see: Privacy International. 'Biometrics: Friend or Foe?', 2017, https://privacyinternational.org/sites/default/files/2017-11/Biometrics_Friend_or_foe.pdf

Introduction

2 K. Hill, 'The Secretive Company that Might End Privacy as We Know It', *The New York Times,* 19 January 2020, https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

3 K. Hill, 'The Secretive Company that Might End Privacy as We Know It', *The New York Times,* 19 January 2020, https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html. Also see: Amba Kak (ed.), 'Regulating Biometrics: Global Approaches and Urgent Questions', AI Now Institute, 1 September 2020, https://ainowinstitute.org/regulatingbiometrics.pdf

4 For instance, controversy around facial recognition in 2020, which culminated in Big Tech backing away from the development and sale of these technologies to varying degrees, did little to scale back facial recognition's public footprint. See e.g.: N. Jansen Reventlow, 'How Amazon's Moratorium on Facial Recognition Tech is Different from IBM's and Microsoft's', *Slate*, 11 June 2020, https://slate.com/technology/2020/06/ibm-microsoft-amazon-facial-recognition-technology.html

5 R.W. Picard, 'Affective Computing', *MIT Media Laboratory Perceptual Computing Section Technical Report*, no. 321, 1995, https://affect.media.mit.edu/pdfs/95.picard.pdf

6 *Market Watch*, Emotion Detection and Recognition (EDR) Market Insights, Status, Latest Amendments and Outlook 2019–2025, 17 September 2020, https://www.marketwatch.com/press-release/emotion-detection-and-recognition-edr-market-insights-status-latest-amendments-and-outlook-2019-2025-2020-09-17.

7 See e.g.: F. Hamilton, 'Police Facial Recognition Robot Identifies Anger and Distress', *The Sunday Times*, 15 August 2020; S. Cha, '"Smile with Your Eyes": How To Beat South Korea's AI Hiring Bots and Land a Job', *Reuters*, 13 January 2020, https://in.reuters.com/article/us-southkorea-artificial-intelligence-jo/smile-with-your-eyes-how-to-beat-south-koreas-ai-hiring-bots-and-land-a-job-idINKBN1ZC022; R. Metz, 'There's a New Obstacle to Landing a Job After College: Getting Approved by AI', *CNN*, 15 January 2020, https://edition.cnn.com/2020/01/15/tech/ai-job-interview/index.html; A. Chen and K. Hao, 'Emotion AI Researchers Say Overblown Claims Give Their Work a Bad Name', *MIT Technology Review*, 14 February 2020, https://www.technologyreview.com/2020/02/14/844765/ai-emotion-recognition-affective-computing-hirevue-regulation-ethics/; D. Harwell, 'A Face-Scanning Algorithm Increasingly Decides Whether You Deserve the Job', *The Washington Post*, 6 November 2019, https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/; *iBorderCtrl*, 'IBorderControl: The Project', 2016, https://www.iborderctrl.eu/The-project; C. Rajgopal, 'SA Startup Launches Facial Recognition Software that Analyses Moods', *The Independent Online*, 29 July 2019, https:/www.iol.co.za/technology/sa-startup-launches-facial-recognition-software-that-analyses-moods-30031112

8 Please see the "Background to Emotion Recognition" section in this report for a detailed analysis of this.

9 It is important to note that China is not the only country where emotion-recognition technology is being developed and deployed. For a comparative overview of emotion- and affect-recognition technology developments in the EU, US, and China, see: S. Krier, 'Facing Affect Recognition', in Asia Society, *Exploring AI Issues Across the United States and*

China series, 18 September 2020, https://asiasociety.org/sites/default/files/inline-files/Affect%20Final.pdf.

10 S.C. Greitens, 'Dealing with Demand For China's Global Surveillance Exports', Brookings Institution, April 2020, https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200428_china_surveillance_greitens_v3.pdf.

11 Y. Yang and M. Murgia, 'Facial Recognition: How China Cornered the Surveillance Market', *The Financial Times*, 7 December 2019, https://www.ft.com/content/6f1a8f48-1813-11ea-9ee4-11f260415385

12 L. Fridman, 'Rosalind Picard: Affective Computing, Emotion, Privacy, and Health', *MIT Media Lab Artificial Intelligence* [podcast], 17 June 2019, https://www.media.mit.edu/articles/rosalind-picard-affective-computing-emotion-privacy-and-health-artificial-intelligence-podcast/

13 A. Gross, M. Murgia, and Y. Yang, 'Chinese Tech Groups Shaping UN Facial Recognition Standards', *Financial Times*, 1 December 2019, https://www.ft.com/content/c3555a3c-0d3e-11ea-b2d6-9bf4d1957a67; C. Juan, 'SenseTime, Tencent, Other AI Firms to Draw Up China's Facial Recognition Standard', *Yicai Global*, 11 December 2019, https://www.yicaiglobal.com/news/sensetime-ant-financial-tencent-team-to-set-up-china-facial-recognition-tenets

14 J. Ding, P. Triolo, and S. Sacks, 'Chinese Interests Take a Big Seat at the AI Governance Table', *DigiChina*, 20 June 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinese-interests-take-big-seat-ai-governance-table/

15 Y. Yang and M. Murgia, 'Facial Recognition: How China Cornered the Surveillance Market', *The Financial Times*, 7 December 2019, https://www.ft.com/content/6f1a8f48-1813-11ea-9ee4-11f260415385

16 For an explainer of AI, see: ARTICLE 19 and Privacy International, *Privacy and Freedom of Expression in the Age of Artificial Intelligence*, 2018, https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf

17 For an analysis of the fundamental nature of facial recognition, see: L. Stark, 'Facial Recognition is the Plutonium of AI', *ACM XRDS*, vol. 25, no. 3, Spring 2019, pp. 50–55. Also see: V. Marda, 'Facial Recognition is an Invasive and Inefficient Tool', *The Hindu*, 22 July 2019, https://www.thehindu.com/opinion/op-ed/facial-recognition-is-an-invasive-and-inefficient-tool/article28629051.ece

18 The article demarcates the first wave of biometric technologies as those that gathered data such as iris scans, voiceprints, palm- and fingerprints, and facial images captured via cameras, as well as temperature and ultrasonic sensors, while biometrics 2.0 incorporated gait analysis. '生物识别3.0时代, 阿尔法鹰眼想用"情感计算"布局智慧安全' ['In the Age of Biometrics 3.0, Alpha Hawkeye Wants to Use "Affective Computing" to Deploy Smart Security'], *Sohu*, 28 April 2017, https://www.sohu.com/a/137016839_114778.; or '多维度识别情绪，这家公司要打造审讯问询的AlphaGo' ['Multimodal Emotion Recognition, This Company Wants to Create the AlphaGo of Interrogation'], *Sohu*, 23 March 2019, https://www.sohu.com/a/303378512_115035

19 '云思创智“灵视多模态情绪研判审讯系统”亮相南京安博会' ['Xinktech "Lingshi Multimodal Emotion Research and Interrogation System" Appeared at Nanjing Security Expo'], *Sohu*, 21 March 2019, https://www.sohu.com/a/302906390_120049574

20 See, in particular: P. Ekman, E. Richard Sorenson, and W.V. Friesen, 'Pan-Cultural Elements in Facial Displays of Emotion', *Science*, vol. 164, no. 3875, 1969, pp. 86–88, https://science.sciencemag.org/content/164/3875/86; P. Ekman, 'Universal Facial Expressions of Emotions'. *California Mental Health Research Digest*, vol. 8, no. 4, 1970, pp. 151–158, https://1ammce38p-kj41n8xkp1iocwe-wpengine.netdna-ssl.com/wp-content/uploads/2013/07/

Universal-Facial-Expressions-of-Emotions1.pdf; P. Ekman, 'Universals and Cultural Differences in Facial Expressions of Emotions', in J. Cole (ed.), *Nebraska Symposium on Motivation*, Lincoln, NB: University of Nebraska Press, 1972, pp. 207–282), https://1am-mce38pkj41n8xkp1iocwe-wpengine.netdna-ssl.com/wp-content/uploads/2013/07/Universals-And-Cultural-Differences-In-Facial-Expressions-Of.pdf.

21 See: P. Ekman and W.V. Friesen, 'Nonverbal Leakage and Clues to Deception', *Psychiatry*, vol. 32, 1969, pp. 88–105, also available from https://www.paulekman.com/wp-content/uploads/2013/07/Nonverbal-Leakage-And-Clues-To-Deception.pdf. Also see: Paul Ekman Group, *What Are Micro Expressions?*, https://www.paulekman.com/resources/micro-expressions/

22 A.L. Hoffman and L. Stark, 'Hard Feelings – Inside Out, Silicon Valley, and Why Technologizing Emotion and Memory Is a Dangerous Idea', *Los Angeles Review of Books*, 11 September 2015, https://lareviewofbooks.org/article/hard-feelings-inside-out-silicon-valley-and-why-technologizing-emotion-and-memory-is-a-dangerous-idea/

23 J.A. Russell, 'Is There Universal Recognition of Emotion from Facial Expression? A Review of the Cross-Cultural Studies', *Psychological Bulletin*, vol. 115, no. 1, 1994, pp. 102–141, https://doi.org/10.1037/0033-2909.115.1.102

24 L.F. Barrett *et al.*, 'Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements', *Psychological Science in the Public Interest*, vol. 20, no. 1, 2019, https://journals.sagepub.com/doi/10.1177/1529100619832930

25 J.A. Russell and J.M. Fernández-Dols, 'Coherence between Emotions and Facial Expressions', *The Science of Facial Expression*, Oxford Scholarship Online, 2017, doi: 10.1093/acprof:oso/9780190613501.001.0001.

26 See e.g.: L.F. Barrett, 'What Faces Can't Tell Us', *The New York Times*, 28 February 2014, https://www.nytimes.com/2014/03/02/opinion/sunday/what-faces-cant-tell-us.html

27 L.F. Barrett, 'Are Emotions Natural Kinds?', *Perspectives on Psychological Science*, vol. 1, no. 1, 2006, doi:10.1111/j.1745-6916.2006.00003.x

28 A. Korte, 'Facial Recognition Technology Cannot Read Emotions, Scientists Say', *American Association for the Advancement of Science*, 16 February 2020, https://www.aaas.org/news/facial-recognition-technology-cannot-read-emotions-scientists-say. Also see: S. Porter, 'Secrets and Lies: Involuntary Leakage in Deceptive Facial Expressions as a Function of Emotional Intensity', *Journal of Nonverbal Behavior*, vol. 36, no. 1, March 2012, pp. 23–37, doi: 10.1007/s10919-011-0120-7

29 See e.g.: M. Price, 'Facial Expressions – Including Fear – May Not Be As Universal As We Thought', *Science*, 17 October 2016, https://www.sciencemag.org/news/2016/10/facial-expressions-including-fear-may-not-be-universal-we-thought; or C. Crivelli, J.A. Russell, S. Jarillo, and J.-M. Fernandez-Dols, 2016. 'The Fear Gasping Face as a Thread Display in a Melanesian Society', *Proceedings of the National Academy of Sciences of the United States of America*, 2016, https://doi.org/10.1073/pnas.1611622113

30 C. Chen *et al.*, 'Distinct Facial Expressions Represent Pain and Pleasure Across Cultures', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 43, 2018, pp. E10013–E10021, https://www.pnas.org/content/115/43/E10013

31 L. Stark, 'Facial Recognition, Emotion and Race in Animated Social Media', *First Monday*, vol. 23, no. 9, 3 September 2018; L. Rhue, 'Racial Influence on Automated Perceptions of Emotions', *SSRN*, November 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765; L. Rhue, 'Emotion-Reading Tech Fails the Racial Bias Test', *The Conversation*, 3 January 2019, https://theconversation.com/emotion-reading-tech-fails-the-racial-bias-test-108404. The company referred to in Professor Rhue's paper, Megvii, is one of the companies sanctioned in the US for supplying authorities in Xinjiang province with face-recognition cameras, used to monitor Uighur citizens.

32 Some recent examples include: L. Safra, C. Chevallier, J. Grezes, and N. Baumard, 'Tracking Historical Changes in Trustworthiness Using Machine Learning Analyses of Facial Cues in Paintings', *Nature Communications*, vol. 11, no. 4728, 2020, https://www.nature.com/articles/s41467-020-18566-7. Also see: Coalition for Critical Technology. 'Abolish the #TechToPrisonTimeline', *Medium*, 23 June 2020, https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16

33 In a similar vein, read about physiognomy's enduring legacy: A. Daub, 'The Return of the Face', *Longreads,* 3 October 2018, https://longreads.com/2018/10/03/the-return-of-the-face/

2. Use Cases

34 S.C. Greitens, 'Dealing With Demand For China's Global Surveillance Exports', *Brookings Institution*, April 2020, https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200428_china_surveillance_greitens_v3.pdf

35 A police mobile phone application, Integrated Joint Operations Platform, logs personal dossiers on Uighurs, which include behavioural observations such as irregular electricity usage (associated with bomb production) or the 'anti-social behaviour' of leaving one's home through a back entrance. M. Wang, *China's Algorithms of Repression: Reverse Engineering a Xinjiang Police Mass Surveillance App*, Human Rights Watch, 1 May 2019, https://www.hrw.org/report/2019/05/02/chinas-algorithms-repression/reverse-engineering-xinjiang-police-mass

36 While the Chinese name 雪亮 (*xuěliàng*) literally translates to 'dazzling snow', this is often used as a figurative expression to describe 'sharp eyes'. See: D. Bandurski, '"Project Dazzling Snow": How China's Total Surveillance Experiment Will Cover the Country', *Hong Kong Free Press*, 12 August 2018, https://hongkongfp.com/2018/08/12/project-dazzling-snow-chinas-total-surveillance-experiment-set-expand-across-country/; D. Peterson and J. Rudolph, 'Sharper Eyes: From Shandong to Xinjiang (Part 3)', *China Digital Times*, 13 September 2019, https://chinadigitaltimes.net/2019/09/sharper-eyes-shandong-to-xinjiang-part-3/; S. Denyer, 'China's Watchful Eye: Beijing Bets on Facial Recognition in a Big Drive for Total Surveillance', *Washington Post*, 7 January 2018, https://www.washingtonpost.com/news/world/wp/2018/01/07/feature/in-china-facial-recognition-is-sharp-end-of-a-drive-for-total-surveillance/

37 Axis Communications, a Swedish company, supplied surveillance cameras to the Sharp Eyes project. See: Amnesty International, *Out of Control: Failing EU Laws for Digital Surveillance Export*, 21 September 2020, https://www.amnesty.org/en/documents/EUR01/2556/2020/en/

38 D. Bandurski, 'A Rock Star's "Fengqiao Experience"', *China Media Project*, 30 November 2018, https://chinamediaproject.org/2018/11/30/chen-yufans-fengqiao-experience/

39 J. Page and E. Dou, 'In Sign of Resistance, Chinese Balk at Using Apps to Snitch on Neighbors', *Wall Street Journal*, 29 December 2018, https://www.wsj.com/amp/articles/in-sign-of-resistance-chinese-balk-at-using-apps-to-snitch-on-neighbors-1514566110. For details on police involvement in Fengqiao-inspired neighbourhood-watch groups, see: N. Gan, 'Police Embedded in Grass-Roots Communist Party Cells as Security Grip Tightens on Beijing', *South China Morning Post*, 18 February 2019, https://www.scmp.com/news/china/politics/article/2186545/police-embedded-grass-roots-communist-party-cells-security-grip

40 To read more on Alpha Hawkeye's participation in Sharp Eyes, see: '原创|人工智能 情感计算反恐缉私禁毒应用新方向' ['Innovation| AI New Directions in Affective Computing Counterterror, Anti-Smuggling, Anti-Drug Applications'], 中国安防行业网 [*China Security Industry Network*], 17 July 2019, http://news.21csp.com.cn/C19/201907/11383210.html; '阿尔法鹰眼' ['Alpha Hawkeye'], Homepage, http://www.alphaeye.com.cn/. For more on ZNV Liwei's

contribution to Sharp Eyes, see: *ZNV Liwei*, '雪亮工程' 解决方案' ['Sharp Eyes Project' Solution'], http://www.znv.com.cn/Anli/anliDetails.aspx?id=100000507363375&no-decode=101003003001. For the Xinktech-Fengqiao link, see: Xinktech, '云思助力'枫桥式公安派出所'建设，AI多模态智能审讯桌首次亮相' ['Xinktech Helps Establish "Fengqiao-Style PSB", Debuts AI Multimodal Smart Interrogation Table'], *WeChat*, 31 March 2020, https://mp.weixin.qq.com/s/n1lIEKWmpsbizhKIgZzXCw

41 Hikvision cameras are one example. See: Z. Shen *et al*., 'Emotion Recognition Based on Multi-View Body Gestures', *ICIP*, 2019, https://ieeexplore.ieee.org/abstract/document/8803460

42 Xinktech, '江苏省公安厅举办刑侦部门审讯专业人才培训 云思创智应邀分享微表情技术实战应用' ['Jiangsu Province's Public Security Department Criminal Investigation Bureau Held Training for Interrogation Professionals, Xinktech Was Invited to Share Practical Applications of Micro Expression Technology'], 29 December 2018, http://www.xinktech.com/news_detail8.html

43 *Sohu*, '想撒谎？不存在的！情绪分析近千亿蓝海市场待挖掘' ['Thinking of Lying? Can't Happen! Emotion Analysis an Almost 100 Billion RMB Untapped Blue Ocean Market'], 23 March 2019, https://www.sohu.com/a/303263320_545428

44 蔡村、陈正东、沈蓓蓓. [C. Cun, C. Zhengdong, and S. Beibei], '把握瞬间真实:海关旅检应用微表情心理学的构想' ['Grasp the Truth in an Instant: Application of Micro-Expressions Psychology in Customs Inspection of Passengers'], 《海关与经贸研究》[*Journal of Customs and Trade*], no. 3, 2018, pp. 29–30.

45 Government Accountability Office, 'Aviation Security: TSA Should Limit Future Funding for Behavior Detection Activities', 14 November 2013, https://www.gao.gov/products/gao-14-158t; Department of Homeland Security Office of Inspector General, 'TSA's Screening of Passengers by Observation Techniques', May 2013, https://www.oig.dhs.gov/assets/Mgmt/2013/OIG_SLP_13-91_May13.pdf; American Civil Liberties Union, *ACLU vs. TSA*, 8 February 2017, https://www.aclu.org/cases/aclu-v-tsa; N. Anderson, 'TSA's Got 94 Signs to ID Terrorists, But They're Unproven by Science', *Ars Technica*, 13 November 2013, https://arstechnica.com/tech-policy/2013/11/despite-lack-of-science-tsa-spent-millions-on-behavioral-detection-officers/; J. Winter and C. Currier, 'Exclusive: TSA's Secret Behavior Checklist to Spot Terrorists', *The Intercept*, 27 March 2015, https://theintercept.com/2015/03/27/revealed-tsas-closely-held-behavior-checklist-spot-terrorists/; J. Krywko, 'The Premature Quest for AI-Powered Facial Recognition to Simplify Screening', *Ars Technica*, 2 June 2017, https://arstechnica.com/information-technology/2017/06/security-obsessed-wait-but-can-ai-learn-to-spot-the-face-of-a-liar/

46 M.S. Shmidt and E. Lichtblau, 'Racial Profiling Rife at Airport, US Officers Say', *The New York Times,* 11 August 2012, https://www.nytimes.com/2012/08/12/us/racial-profiling-at-boston-airport-officials-say.html

47 J. Sánchez-Monederoa and L. Dencik, 'The Politics of Deceptive Borders: "Biomarkers of Deceit" and the Case of iBorderCtrl', Cardiff, School of Journalism, Media and Cullture, Cardiff University, n.d., https://arxiv.org/ftp/arxiv/papers/1911/1911.09156.pdf; R. Gallagher and L. Jona, 'We Tested Europe's New Lie Detector for Travelers – and Immediately Triggered a False Positive', *The Intercept*, 26 July 2019, https://theintercept.com/2019/07/26/europe-border-control-ai-lie-detector/

48 *Sohu*, 'iRank: 基于互联网类脑架构的阿尔法鹰眼发展趋势评估' ['iRank: Analysis of Alpha Hawkeye's Internet-like Brain Architecture Development Trend'], 28 March 2018, https://www.sohu.com/a/226632874_297710

49 The language used to describe early-warning systems in China is reminiscent of how foreign emotion recognition firms that sought applications in forensics and judicial procedures, like Cephos and No Lie MRI, first pitched their (seemingly now defunct) platforms. For Cephos, see: *Business Wire*, 'Cephos Corporation to Offer Breakthrough Deception Detection

Services Using fMRI Technology with over 90% Accuracy in Individuals', 27 September 2005, https://www.businesswire.com/news/home/20050927005072/en/Cephos-Corporation-Offer-Breakthrough-Deception-Detection-Services and Cephos official company website, http://www.cephoscorp.com/about/. For No Lie MRI, see: A. Madrigal, 'MRI Lie Detection to Get First Day in Court', *WIRED*, 13 March 2009, https://www.wired.com/2009/03/noliemri/; J. Calderone, 'There Are Some Big Problems with Brain-Scan Lie Detectors', *Business Insider*, 19 April 2016, https://www.businessinsider.com/dr-oz-huizenga-fmri-brain-lie-detector-2016-4

50 段蓓玲 [Duan Beiling], '视频侦查主动预警系统应用研究' ['Applied Research on Active Early Warning System for Video Investigations'], 《法制博览》[*Legality Vision*], no. 16, 2019. This paper was funded as part of the 2017 Hubei Province Department of Education's Youth Talent 'Research on Active Video Investigation in the Context of Big Data' project.

51 *Ibid.*

52 *Ibid.*

53 刘缘、庾永波 [L. Yan and L. Yongbo],'在安检中加强"微表情"识别的思考 – – 基于入藏公路安检的考察'['Strengthening the Application of Micro-Expression in Security Inspection: Taking the Road Security Check Entering Tibet as an Example'], 《四川警察学院学报》[*Journal of Sichuan Police College*], vol. 30, no. 1, February 2019.

54 *Ibid.*

55 *Ibid.*

56 Early descriptions of 'Alpha Eye' closely mirror later write-ups of Alpha Hawkeye, which itself has translated its name as 'Alpha Eye' in some images. This report assumes both names refer to the same company. 察网 [*Cha Wang*], '比"阿法狗"更厉害的是中国的"阿尔法眼"' ['China's "Alpha Eye" is More Powerful Than an "Alpha Dog"'], 17 March 2016, http://www.cwzg.cn/politics/201603/26982.html; 杨丽 [Y. Li], "阿尔法眼'义乌试验两天 查到5个带两张身份证的人' ['"Alpha Eye" Trialled in Yiwu for Two Days, Finds 5 People Carrying Two State ID Cards'], China.com.cn, 31 March 2016, http://zjnews.china.com.cn/zj/hz/2016-03-31/68428.html

57 中国安防行业网 [*China Security Industry Network*], '原创|人工智能 情感计算反恐缉私禁毒应用新方向' ['Innovation | AI New Directions in Affective Computing Counterterror, Anti-smuggling, Anti-drug Applications'], 17 July 2019, http://news.21csp.com.cn/C19/201907/11383210.html; '阿尔法鹰眼' ['Alpha Hawkeye'], homepage, http://www.alphaeye.com.cn/

58 ZNV Liwei, '"雪亮工程"解决方案' ['"Sharp Eyes Project" Solution'] http://www.znv.com.cn/Anli/anliDetails.aspx?id=100000507363375&nodecode=101003003001; ZNV Liwei, 'ZNV力维与南京森林警察学院成立'AI情绪大数据联合实验室'!' ['ZNV Liwei and Nanjing Forest Police College Establish "AI Emotion Big Data Joint Lab", http://www.znv.com.cn/About/NewsDetais.aspx?id=100000615683266&nodecode=101006002001

59 '生物识别3.0时代, 阿尔法鹰眼想用"情感计算"布局智慧安全' ['In the Age of Biometrics 3.0, Alpha Hawkeye wants to use "Affective Computing" to Deploy Smart Security'], *Sohu*, 28 April, 2017. https://www.sohu.com/a/137016839_114778

60 *Sohu*, '生物识别3.0时代，阿尔法鹰眼想用"情感计算"布局智慧安全' ['In the Age of Biometrics 3.0, Alpha Hawkeye wants to use "Affective Computing" to Deploy Smart Security'], 28 April 2017, https://www.sohu.com/a/137016839_114778; 《南京日报》[*Nanjing Daily*], '多个城市已利用AI读心加强反恐安防' ['Several Cities Have Used AI Mind Reading to Strengthen Counterterrorist Security'], 29 September 2018, http://njrb.njdaily.cn/njrb/html/2018-09/29/content_514652.htm; *Sohu*, 'iRank: 基于互联网类脑架构的阿尔法鹰眼发展趋势评估' ['iRank: Analysis of Alpha Hawkeye's Internet-like Brain Architecture Development Trend'], 28 March 2018, https://www.sohu.com/a/226632874_297710; 云涌 [*Yunyong (Ningbo News)*], '专访之三:看一眼就读懂你,甬企这双"鹰眼"安防科技够 "黑" ['Interview 3: It Can Understand You in One Glance, This Ningbo Company's Pair of "Hawk Eyes"

Security Technology is "Black" Enough'], 4 May 2018, http://yy.cnnb.com.cn/system/2018/05/04/008748677.shtml

61 CM Cross, '卡口应用' ['Customs Applications'], 23 November 2018, http://www.cmcross.com.cn/index.php/Home/List/detail/list_id/104

62 CM Cross, '侦讯应用' ['Interrogation Applications'], 9 November 2018, http://www.cmcross.com.cn/index.php/Home/List/detail/list_id/106

63 Taigusys Computing, '太古计算行为分析技术让生活更智能，更美好!' ['Taigusys Computing's Behavioral Analysis Technology Makes Life Smarter, More Beautiful!'], 11 April 2019, http://www.taigusys.com/news/news145.html; Taigusys Computing, '产品中心' ['Product Center'], http://www.taigusys.com/procen/procen139.html

64 Other than medical- and financial-services-related applications of the technology, all suggested and already implemented uses in Table 1 are undertaken by government and law enforcement institutions, including police, public security bureaux, customs and border inspection agencies, and prisons.

65 云涌 [*Yunyong (Ningbo News)*], '专访之三：看一眼就读懂你, 甬企这双 "鹰眼" 安防科技够 "黑" ['Interview 3: It Can Understand You in One Glance, This Ningbo Company's Pair of "Hawk Eyes" Security Technology is "Black" Enough'], 4 May 2018, http://yy.cnnb.com.cn/system/2018/05/04/008748677.shtml; 《南京日报》[*Nanjing Daily*], '人工智能"鹰眼"如何看穿人心？释疑 它和"微表情", 测谎仪有何不同' ['How Does AI "Eagle Eye" See Through People's Hearts? Explanation of How it Differs From "Microexpressions" and Polygraph'], 29 September 2018, http://njrb.njdaily.cn/njrb/html/2018-09/29/content_514653.html

66 CM Cross, '公司简介' ['Company Profile'], 9 November 2018, http://www.cmcross.com.cn/index.php/Home/List/detail/list_id/151

67 CM Cross, '卡口应用' ['Customs Applications'], 23 November 2018. http://www.cmcross.com.cn/index.php/Home/List/detail/list_id/104

68 BOSS 直聘 [BOSS Zhipin], 'EmoKit简介' ['Brief Introduction to EmoKit'], https://www.zhipin.com/gongsi/6a438b988fa2bb8003x63d6_.html

69 中国青年网 [Youth.cn], "曲靖模式"先行项目 – – 翼开科技，成功在曲落地扎根' ['The First "Qujing Style" Project-EmoKit Technology Successfully Takes Root in Quluo'], 5 September 2019, http://finance.youth.cn/finance_cyxfgsxw/201909/t20190905_12062221.htm

70 爱分析 [*ifenxi*], '翼开科技CEO魏清晨：金融反欺诈是AI多模态情感计算最佳落地场景' ['EmoKit CEO Wei Qingchen: Finance Anti-Fraud is AI Multimodal Affective Computing's Best Landing Scenario'], 29 August 2018, https://ifenxi.com/research/content/4164

71 DeepAffex, 'Security Use Case', https://www.deepaffex.ai/security; T. Shihua, 'China's Joyware Electronics Joins Hands with NuraLogix on Emotional AI', *Yicai Global*, 25 October 2018, https://www.yicaiglobal.com/news/china-joyware-electronics-joins-hands-with-nuralogix-on-emotional-ai

72 T. Shihua, 'Joyware Bags Sole China Rights to Nuralogix's Emotion-Detecting AI Tech', *Yicai Global*, 5 November 2019, https://www.yicaiglobal.com/news/joyware-bags-sole-china-rights-to-nuralogix-emotion-detecting-ai-tech

73 *Sohu*, '他们举起摄像头 3秒扫描面部测心率 秒懂你情绪' ['They Held the Camera Up and Scanned Their Faces for 3 Seconds to Measure Their Heart Rate, Understand Your Emotions in Seconds'], 24 September 2016, https://www.sohu.com/a/114988779_270614

74 Sage Data, '公安多模态情绪审讯系统' ['Public Security Multimodal Emotion Interrogation System'], http://www.sagedata.cn/#/product/police

75 Anshibao, '反恐治安： 情绪识别系统' ['Counterterror Law and Order: Emotion Recognition System'], http://www.asbdefe.com/cn/product_detail-925080-1933004-176891.html

76 警用装备网 [Police Equipment Network], '产品名称：动态情绪识别' ['Product Name: Dynamic Emotion Recognition'], http://www.cpspew.com/product/987395.html.

77 Taigusys, '太古计算行为分析技术，让生活更智能，更美好!' ['Taigusys Computing's Behavioral Analysis Technology Makes Life Smarter, More Beautiful!'], 11 April 2019, http://www.taigusys.com/news/news145.html

78 Taigusys, '产品中心' ['Product Center'], January, http://www.taigusys.com/procen/procen139.html

79 Taigusys, 公安部门的摄像头每天都拍到了什么？答案令人吃惊' ['What Do Public Security Bureaux' Cameras Record Every Day? The Answer is Shocking!'], 5 January 2019, http://www.taigusys.com/news/news115.html

80 Xinktech, '云思创智"灵视 – – 多模态情绪审讯系统"亮相"新时代刑事辩护"高端论坛' ['Xinktech's "Lingshi-Multimodal Emotional Interrogation System" Featured in "New Era Criminal Defence" High-end Forum'], http://www.xinktech.com/news_detail10.html; Xinktech, '情绪识别' ['Emotion Recognition'], http://www.xinktech.com/emotionRecognition.html

81 Xinktech, '云思创智'灵视 – – 多模态情绪审讯系统'亮相'新时代刑事辩护'高端论坛' ['Xinktech's 'Lingshi-Multimodal Emotional Interrogation System' Featured in "New Era Criminal Defence" High-end Forum'], http://www.xinktech.com/news_detail10.html; Xinktech, '江苏省公安厅举办刑侦部门审讯专业人才培训 云思创智应邀分享微表情技术实战应用' ['Jiangsu Province's Public Security Department Criminal Investigation Bureau Held Training for Interrogation Professionals, Xinktech Was Invited to Share Practical Applications of Micro Expression Technology'], 29 December 2018, http://www.xinktech.com/news_detail8.html

82 Xinktech, '云思助力'枫桥式公安派出所'建设，AI多模态智能审讯桌首次亮相' ['Xinktech Helps Establish "Fengqiao-Style PSB", Debuts AI Multimodal Smart Interrogation Table'], *WeChat*, 31 March 2020, https://mp.weixin.qq.com/s/n1lIEKWmpsbizhKIgZzXCw

83 *Sohu*, '想撒谎？不存在的！情绪分析近千亿蓝海市场待挖掘' ['Thinking of Lying? Can't Happen! Emotion Analysis an Almost 100 Billion RMB Untapped Blue Ocean Market'], 23 March 2019, https://www.sohu.com/a/303263320_545428

84 华强智慧网 [Huaqiang Smart Network], 'ZNV力维推出心理与情绪识别系统 主要适用于公安审讯场景' ['ZNV Liwei Launches Psychological and Emotional Recognition System, Mainly Used in Public Security Interrogation Settings'], 11 April 2019, http://news.hqps.com/article/201904/301837.html

85 *Huanqiu*, 'AI赋能·智享未' 灵视多模态情绪研判系统产品发布会在南京隆重召开' ['"AI Empowerment · Enjoy the Future" Lingshi Multimodal Emotion Research and Judgment System Product Launch Conference Held in Nanjing.'], 25 March 2019, https://tech.huanqiu.com/article/9CaKrnKjhJ7; *Sohu*, '多维度识别情绪，这家公司要打造审讯问询的AlphaGo' ['Multimodal Emotion Recognition, This Company Wants to Create the AlphaGo of Interrogation'], 23 March 2019, https://www.sohu.com/a/303378512_115035. AlphaGo is the DeepMind-developed computer program that used neural networks to defeat the world champion of Go, a complex, millennia-old Chinese board game of strategy. See: DeepMind, 'AlphaGo', https://deepmind.com/research/case-studies/alphago-the-story-so-far

86 Xinktech, '云思创智"灵视 – – 多模态情绪审讯系统"亮相 "新时代刑事辩护"高端论坛' ['Xinktech's "Lingshi-Multimodal Emotional Interrogation System" Featured in "New Era Criminal Defence" High-end Forum'], http://www.xinktech.com/news_detail10.html. In a campy promotional video, Xinktech superimposed the platform's interface on footage from a popular Chinese TV drama; as a handcuffed detainee in an orange jumpsuit replies to investigators' questions, his face is encased in a red box, alongside text annotating his emotional state and heart rate. Within the dashboard containing the video feed is a series of graphs measuring real-time emotional and physiological responses. Xinktech, '云思创智公安多模态情绪审讯系统及应用场景' ['Xinktech Public Security Multimodal Emotion Interrogation System and Application Scenarios'], 23 October 2018, http://www.xinktech.com/news_detail1.html

87 T. Shihua, 'China's Joyware Electronics Joins Hands with NuraLogix on Emotional AI', *Yicai Global*, 25 October 2018, https://www.yicaiglobal.com/news/china-joyware-electronics-joins-hand-swith-nuralogix-on-emotional-ai

88 A. Al-Heeti. 'This App Could Turn Your Phone into a Lie Detector', *CNET*, 20 April 2018, https://www.cnet.com/news/your-phone-could-become-a-tool-for-detecting-lies-with-veritaps/; K. Fernandez-Blance, 'Could Your Face be a Window to Your Health? U of T Startup Gathers Vital Signs from Video', *University of Toronto News*, 24 March 2017, https://www.utoronto.ca/news/could-your-face-be-window-your-health-u-t-startup-gathers-vital-signs-video

89 ZNV '力维与南京森林警察学院成立"AI情绪大数据联合实验室"!' ['ZNV Liwei and Nanjing Forest Police College Establish "AI Emotion Big Data Joint Lab"!']. ZNV Liwei official company website. http://www.znv.com.cn/About/NewsDetais.aspx?id=100000615683266&no-decode=101006002001

90 邓翠平[D. Cuiping].'中南财经政法大学刑事司法学院创建人工智能联合实验室' ['Zhongnan University of Economics and Law Establishes AI Joint Lab'], *Xinhua*, 16 March 2019, http://m.xinhuanet.com/hb/2019-03/16/c_1124242401.htm; 周苏展 [Z. Suzhan], 研究"读心术,"中南大人工智能联合实验室揭牌" ['Researching "Mind Reading Technique", Zhongnan University of Economics and Law Unveils AI Joint Lab'], *Zhongnan University of Economics and Law Alumni News*, 18 March 2019, http://alumni.zuel.edu.cn/2019/0318/c415a211961/pages.htm. Additional signatories of the cooperative agreement with People's Public Security University include: Interrogation Science and Technology Research Center of the People's Public Security University of China, School of Criminal Justice of Zhongnan University of Economics and Law, Beijing AVIC Antong Technology Co. Ltd., Hebei Hangpu Electronics Technology Co. Ltd., Tianjin Youmeng Technology Co. Ltd., and Nanjing Qingdun Information Technology Co. Ltd. (中南财经政法大学刑事司法学院、北京中航安通科技有限公司、河北航普电子科技有限公司、天津友盟科技有限公司、南京擎盾信息科技有限公司签署了战略合作), see: *Huanqiu*, "AI赋能·智享未" 灵视多模态情绪研判系统产品发布会在南京隆重召开' ['"AI Empowerment · Enjoy the Future" Lingshi Multimodal Emotion Research and Judgment System Product Launch Conference Held in Nanjing'], 25 March 2019, https://tech.huanqiu.com/article/9CaKrnKjhJ7

91 *Sohu*, '想撒谎？不存在的！情绪分析近千亿蓝海市场待挖掘' ['Thinking of Lying? Can't Happen! Emotion Analysis an Almost 100 Billion RMB Untapped Blue Ocean Market'], 23 March 2019, https://www.sohu.com/a/303263320_545428. Xinktech's official WeChat account lists Hikvision, China Communications Bank, iFlytek, and State Grid among two dozen business partners. A list of academic partnerships includes seven Chinese universities and think tanks, four US universities, and one Canadian university. See: Xinktech, '云思助力"枫桥式公安派出所"建设，AI多模态智能审讯桌首次亮相'，AI多模态智能审讯桌首次亮相' ['Xinktech Helps Establish "Fengqiao- Style PSB", Debuts AI Multimodal Smart Interrogation Table'], *WeChat*, 31 March 2020, https://mp.weixin.qq.com/s/n1lIEKWmpsbizhKIgZzXCw

92 BOSS 直聘 [BOSS Zhipin], 'EmoKit简介' ['Brief Introduction to EmoKit'], https://www.zhipin.com/gongsi/6a438b988fa2bb8003x63d6_.html

93 铅笔道 [*Qianbidao*], '他用AI情感算法来做"测谎仪" 已为网贷公司提供反骗贷服务 获订单300万' ['He Used AI Emotion Algorithms to Make a "Lie Detector", Has Already Provided Anti-fraud Services to Online Lending Companies and Won 3 Million Orders'], 23 July 2018, https://www.pencilnews.cn/p/20170.html

94 *Sohu*, '他们举起摄像头 3秒扫描面部测心率 秒懂你情绪' ['They Held the Camera Up and Scanned Their Faces for 3 Seconds to Measure Their Heart Rate, Understand Your Emotions in Seconds'], 24 September 2016, https://www.sohu.com/a/114988779_270614

95 *Ibid*.

96 *Sohu*, '想撒谎？不存在的！情绪分析近千亿蓝海市场待挖掘' ['Thinking of Lying? Can't Happen! Emotion Analysis an Almost 100 Billion RMB Untapped Blue Ocean Market'], 23 March 2019, https://www.sohu.com/a/303263320_545428. The article further elaborates that training

a model requires at least 10,000 such data samples, with each costing 2,000–3,000 yuan (USD305–457).

97 Xinktech, '强强联合，推动行业进步 – – 云思创智受邀与"中南财经政法大学刑事司法学院"进行技术交流' [Strong Alliance to Promote Industry Progress- – – Xinktech Invited to Conduct a Technical Exchange With "School of Criminal Justice, Zhongnan University of Economics and Law"'], 29 December 2018, http://www.xinktech.com/news_detail9.html

98 *Ibid.*

99 *Sohu*, '云思创智"灵视多模态情绪研判审讯系统"亮相南京安博会' ['Xinktech "Lingshi Multimodal Emotion Research and Interrogation System" Appeared at Nanjing Security Expo'], 21 March 2019, https://www.sohu.com/a/302906390_120049574

100 Xinktech, '江苏省公安厅举办刑侦部门审讯专业人才培训 云思创智应邀分享微表情技术实战应用' ['Jiangsu Province's Public Security Department Criminal Investigation Bureau Held Training for Interrogation Professionals, Xinktech Was Invited to Share Practical Applications of Micro Expression Technology'], 29 December 2018, http://www.xinktech.com/news_detail8.html, and identical article on Xinktech's *WeChat* public account: https://mp.weixin.qq.com/s/InCb8yR68v1FiMOaS-JZwMA. The same year, Xinktech won the Jiangsu Province Top Ten AI Products Award; see: Xinktech, '云思创智"沉思者智能算法建模训练平台"荣获"2018年度江苏省十佳优秀人工智能产品"奖' ['Xinktech's "Thinker Intelligent Algorithm Model-Building Training Platform" Wins "2018 Jiangsu Top Ten Excellent AI Products" Award'], 8 September 2018, http://www.xinktech.com/news_detail3.html

101 中国青年网 [Youth.cn], '"曲靖模式"先行项目 – – 翼开科技，成功在曲落地扎根 ['The First "Qujing Style" Project-EmoKit Technology Successfully Takes Root in Quluo'], 5 September 2019, http://finance.youth.cn/finance_cyxfgsxw/201909/t20190905_12062221.htm

102 赵青晖 [Z. Qinghui], '凭借识别人的情绪，他们做到了2000多万用户、1000多万订单' ['Relying on Recognizing Emotions, They Reached Over 20 Million Users and More Than 10 Million Orders'], 芯基建 ['Core Infrastructure' *WeChat* public account], 1 June 2017, https://mp.weixin.qq.com/s/JdhZbS4Ndb_mfq4dV7A0_g

103 US-based multimodal emotion-recognition provider, Eyeris, featured an in-vehicle product installed in a Tesla and announced interest from UK and Japanese auto manufacturers. *C.f.* e.g. *PR Newswire*, 'Eyeris Introduces World's First In-Cabin Sensor Fusion AI at CES2020', 6 January 2020, https://www.prnewswire.com/news-releases/eyeris-introduces-worlds-first-in-cabin-sensor-fusion-ai-at-ces2020-300981567.html. Boston-based Affectiva is one of at least half a dozen other companies known to be developing in-vehicle emotion-based driver-safety products. Some in the field anticipate that emotion-sensing technologies in cars will become mainstream within the next three years. See M. Elgan, 'What Happens When Cars Get Emotional?', *Fast Company*, 27 June 2019, https://www.fastcompany.com/90368804/emotion-sensing-cars-promise-to-make-our-roads-much-safer. Companies working on driver-fatigue recognition include EyeSight Technologies, Guardian Optical Technologies, Nuance Automotive, Smart Eye, and Seeing Machines. For details on the Euro NCAP program, see: S. O'Hear. 'EyeSight Scores $15M to use computer vision to combat driver distraction', *Tech Crunch*, 23 October 2018, https://techcrunch.com/2018/10/23/eyesight. The EU has sponsored the European New Car Assessment Programme, a voluntary vehicle-safety rating system that calls for inclusion of driver-monitoring systems.

104 *Sohu*, '乐视无人驾驶超级汽车亮相6股有望爆发' ['LeEco Driverless Car VV6 Shares Expected to Blow Up'], 21 April 2016, https://www.sohu.com/a/70619656_115411; M. Phenix, 'From China, A Shot Across Tesla's Bow', BBC, 21 April 2016, http://www.bbc.com/autos/story/20160421-from-china-a-shot-across-teslas-bow

105 Great Wall Motor Company Limited, *2019 Corporate Sustainability Report*, p. 34, https://res.gwm.com.cn/2020/04/26/1657959_130_E-12.pdf

106 See: 长城网 [Hebei.com.cn], '2019年那些用过就无法回头的汽车配置: L2自动驾驶' ['In 2019 There's No Looking Back After Using Those Cars' Configurations: L2 Autonomous Driving'], 3 January 2020, http://auto.hebei.com.cn/system/2020/01/02/100152238.shtml

107 Sina Cars［新浪汽车］, '长城汽车发布生命体征监测技术 2021款WEY VV6将成为首款车型' [Great Wall Motors Releases Vital Signs Monitoring Technology, The 2021 Model of VV6 Will Become the First Model'], 8 June 2020, https://auto.sina.com.cn/newcar/2020-06-08/detail-iircuyvi7378483.shtml

108 重庆晨报 [*Chongqing Morning Post*, '上游新闻直播 "UNI-TECH" 科技日活动，秀智能长安汽车实力圈粉' [Upstream News Broadcasts "UNI-TECH" Technology Day Events, Shows Chang'an Automobile's Power-Fans'], 11 April 2020, https://www.cqcb.com/qiche/2020-04-12/2323984_pc.html

109 *Ibid*

110 *China Daily*, '真车来了！华为 HiCar在卓悦中心展示硬核智慧出行服务' ['A Real Car Has Arrived! Huawei Showcases HiCar's Hard-Core Smart Transportation Services at One Avenue Center'], 8 June 2020, http://cn.chinadaily.com.cn/a/202006/08/WS5eddda77a31027ab2a-8ceed4.html. Startups specialising in various AI applications including voice and emotion recognition have also been known to supply these capabilities to car companies, such as the company AI Speech's (思必驰) partnerships with two state-owned car manufacturers, BAIC Group and FAW Group. See: *Sina Finance*, '华为、英特尔、富士康等合作伙伴 AI企业思必驰完成E轮4.1亿元融资' ['Huawei, Intel, Foxconn, and Other Cooperative Partners of AI Company AISpeech Complete E Round of 410 Million Yuan Financing'], 7 April 2020, https://finance.sina.cn/2020-04-07/detail-iimxxsth4098224.d.html

111 杨雪娇 [Y. Xuejiao], '发力行为识别技术 太古计算的AI生意经' ['Generating Momentum in Behavior Recognition Technology, Taigusys Computing's AI Business Sense'], *CPS*中安网 ['CPS Zhong'an Network' WeChat public account], 24 June 2019, https://mp.weixin.qq.com/s/Q7_Kqghotd7X38qXw4gLCg

112 *Sina Cars* [新浪汽车], '车内生物监测、乘员情绪识别，爱驰U5的黑科技你了解多少?' ['Biometric Monitoring and Passenger Emotion Recognition in Vehicles, How Much Do You Understand About Aichi U5's Black Technology?'], 19 July 2019, https://k.sina.com.cn/article_5260903737_13993053900100iug3.html; AIWAYS, '深化探索AI突围之路 爱驰汽车亮相2019中国国际智能产业博览会' ['AIWAYS Shows Deep Exploration of AI Breakthroughs at 2019 China International Smart Industry Expo'], 27 August 2019, https://www.ai-ways.com/2019/08/27/9162/

113 凤凰网 [*iFeng News*], '打造保险科技转型急先锋 平安产险携多项AI技术亮相世界人工智能大会' ['Creating a Pioneer in the Transformation of Insurance Technology, Ping An Property Insurance Company Showcases Several AI Technologies at the World Artificial Intelligence Conference'], 29 August 2019, http://sn.ifeng.com/a/20190829/7693650_0.shtml

114 *Xinhua*, '金融与科技加速融合迈入"智能金融时代" ['Accelerate the Fusion of Finance and Technology Into the "Era of Smart Finance"'], 30 August 2019, http://www.xinhuanet.com/2019-08/30/c_1124942152.htm

115 姬建岗、 郭晓春、张敏、冯春强l [J. Jiangang *et al.*], '人脸识别技术在高速公路打逃中的应用探讨' ['Discussion on Application of Face Recognition Technology in Highway [Toll] Evasion'],《中国交通信息化》[*China ITS Journal*], no. 1, 2018.

116 *Ibid.*

117 *Ibid.*

118 *Ibid.*

119 广州市科学技术局 [Guangzhou Municipal Science and Technology Bureau], '广州市重点领域研发计划 2019 年度"智能网联汽车"(征求意见稿)' ['Guangzhou Key Areas for Research and Development Annual Plan 2019 "Smart Connected Cars" (Draft for Comments'], pp. 5–6, http://kjj.gz.gov.cn/GZ05/2.2/201908/b6444d5e26fc4a628fd7e90517dff499/files/452599ab52df422c999075acf19a3654.pdf

120 *Huanqiu*, "AI赋能·智享未"灵视多模态情绪研判系统产品发布会在南京隆重召开' ['"AI Empowerment · Enjoy the Future" Lingshi Multimodal Emotion Research and Judgment System Product Launch Conference Held in Nanjing'], 25 March, 2019. https://tech.huanqiu.com/article/9CaKrnKjhJ7; *Sohu*, '云思创智'灵视多模态情绪研判审讯系统'亮相南京安博会' ['Xinktech "Lingshi Multimodal Emotion Research and Interrogation System" Appeared at Nanjing Security Expo'], 21 March 2019, https://www.sohu.com/a/302906390_120049574

121 Xinktech, 云思创智CEO凌志辉先生接受中国财经峰 会独家专访 Yun Zhi Chuangzhi CEO Mr Ling Zhihui accepted an exclusive interview with China Finance Summit: http://www.xinktech.com/news_detail5.html. For context on Didi Chuxing passenger assaults, see: S.-L. Wee, 'Didi Suspends Carpooling Service in China After 2nd Passenger Is Killed', *The New York Times*, 26 August 2018, https://www.nytimes.com/2018/08/26/business/didi-chuxing-murder-rape-women.html

122 Hong Kong University of Science and Technology tested an emotion-recognition system on toddlers in Hong Kong and Japan. Teachers would receive data analytics on individual students' attention and emotion levels, as well as aggregated data on classes, see: E. Waltz, 'Are Your Students Bored? This AI Could Tell You', *IEEE Spectrum*, March 2020, https://spectrum.ieee.org/the-human-os/biomedical/devices/ai-tracks-emotions-in-the-classroom; In 2017, the Ecole Supérieure de Gestion (ESG) business school in Paris applied eye-tracking and facial expression-monitoring software from edtech company, Nestor, to detect attention levels of online class attendees, see: A. Toor, 'This French School is Using Facial Recognition to Find Out When Students Aren't Paying Attention', *The Verge*, 26 May 2017, https://www.theverge.com/2017/5/26/15679806/ai-education-facial-recognition-nestor-france. Companies allegedly using face and gesture recognition to detect aggression have supplied schools in Florida and New York with their products, despite criticisms that these emotions, and ensuing dangerous incidents, are extremely difficult to predict, see: For Florida examples, see: D. Harwell, 'Parkland School Turns to Experimental Surveillance Software that can Flag Students as Threats', *Washington Post*, 13 February 2019, https://www.washingtonpost.com/technology/2019/02/13/parkland-school-turns-experimental-surveillance-software-that-can-flag-students-threats/. For New York example, see: M. Moon, 'Facial Recognition is Coming to US Schools, Starting in New York', *Engadget*, 30 May 2019, https://www.engadget.com/2019-05-30-facial-recognition-us-schools-new-york.html. On the difficulty of predicting aggression and safety risks, see: E. Darragh, 'Here's Why I'm Campaigning Against Facial Recognition in Schools', *Motherboard*, 11 March 2020, https://www.vice.com/en_us/article/z3bgpj/heres-why-im-campaigning-against-facial-recognition-in-schools. Also see: P. Vasanth B.A., 'Face Recognition Attendance System in Two Chennai Schools', *The Hindu*, 14 September 2019, https://www.thehindu.com/news/cities/chennai/face-recognition-attendance-system-in-two-city-schools/article29412485.ece

123 S. Swauger, 'Software That Monitors Students During Tests Perpetuates Inequality and Violates Their Privacy', *MIT Technology Review*, 7 August 2020, https://www.technologyreview.com/2020/08/07/1006132/software-algorithms-proctoring-online-tests-ai-ethics/. In China, cameras have been used to proctor primary and secondary school exams for years: http://edu.sina.com.cn/zxx/2018-07-18/doc-ihfnsvyz9043937.shtml

124 *Ibid.*

125 Y.-L. Liu, 'The Future of the Classroom? China's Experience of AI in Education.' In Nesta, *The AI Powered State: China's Approach to Public Sector Innovation*, 18 May 2020, pp. 27–33, https://media.nesta.org.uk/documents/Nesta_TheAIPoweredState_2020.pdf

126 中华人民共和国教育部 [Ministry of Education of the People's Republic of China], '教育部关于印发《高等学校人工智能创新行动计划》的通知' ['Ministry of Education Publishes Notice Regarding Issuance of "Innovative Action

Plan for Artificial Intelligence in Institutions of Higher Education'"], 3 April 2018, http://www.moe.gov.cn/srcsite/A16/s7062/201804/t20180410_332722.html

127 M. Sheehan, 'How China's Massive AI Plan Actually Works', *MacroPolo*, 12 February 2018, https://macropolo.org/analysis/how-chinas-massive-ai-plan-actually-works/

128 李保宏 [L. Baohong], '人工智能在中俄两国教育领域发展现状及趋势' ['The Status Quo and Trend of Artificial Intelligence in the Field of Education in China and Russia'], *Science Innovation*, vol. 7, no. 4, 2019, p.134, http://sciencepg.org/journal/archive?journalid=180&issueid=1800704

129 J. Knox, 'Artificial Intelligence and Education in China', *Learning, Media and Technology*, vol. 45, no. 3, p. 10, https://doi.org/10.1080/17439884.2020.1754236

130 臧威佳、董德森 [Z. Weijia and D. Desen], '人工智能在基础教育中的应用构想.' ['Concepts for Applying Artificial Intelligence in Basic Education'],《西部素质教育》 [*Western China Quality Education*], no. 6, 2019, p. 130. For Meezao marketing concerning the link between emotion recognition and early childhood development, see: *Tianyancha* [天眼查], '蜜枣网：自主研发情绪智能分析系统，深度改变零售与幼教' ['Meezao: Independent Research and Development of Emotion Intelligence Analytical Systems, Deeply Changing Logistics and Preschool Education'], 28 June 2018, https://news.tianyancha.com/ll_i074g8rnrd.html

131 徐姜琴、张永锋 [X. Jiangqin and Z. Yongfeng], '面向慕课的情绪识别系统' ['A MOOC-Oriented Emotion Recognition System'],《创新教育研究》 [*Creative Education Studies*], vol. 6, no. 4, 2018, pp. 299–305.

132 *Ibid.*

133 曹晓明、张永和、潘萌、 朱姗、闫海亮 [C. Xiaoming *et al.*], '人工智能视域下的学习参与度识别方法研究 – 基于一项多模态数据融合的深度学习实验分析' ['Research on Student Engagement Recognition Method from the Perspective of Artificial Intelligence: Analysis of Deep Learning Experiment based on a Multimodal Data Fusion'],《远程教育杂志》 [*Journal of Distance Education*], no. 1, 2019, pp. 32–44.

134 *Ibid.*

135 贾鹂宇、张朝晖、 赵小燕、闫晓炜 [J. Liyu *et al.*], '基于人工智能视频处理的课堂学生状态分析' ['Analysis of Students Status in Class Based on Artificial Intelligence and Video Processing'],《现代教育技术》[*Modern Educational Technology*], no. 12, 2019, pp. 82–88.

136 向鸿炼 [X. Honglian], '人工智能时代下教师亲和力探究' ['On Teacher Affinity in the Era of Artificial Intelligence'],《软件导刊》 [*Software Guide*], no. 11, 2019, pp. 218–221.

137 D. Bryant and A. Howard, 'A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity', *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, January 2019, pp. 377–382, https://doi-org.stanford.idm.oclc.org/10.1145/3306618.3314284

138 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, No. 3, Spring 2019, p. 8, https://thedisconnect.co/three/camera-above-the-classroom/

139 *Ibid.*

140 *Ibid.*

141 芥末堆 [*Jiemodui*], '英孚少儿英语推出"线上+线下"双翼战略，将携手腾讯开发智慧课堂' ['EF English Launches "Online + Offline" Two-Winged Strategy, Will Collaborate With Tencent to Develop Smart Classrooms'], 27 March 2019, https://www.jiemodui.com/N/105395.html; 腾讯云 [*Tencent Cloud*], '从平等化学习、个性化教学、智慧化校园管理3个维度，助力教育公平发展' [Assisting Equitable Development of Education From Three Dimensions of Equal Learning, Personalized Teaching, and Smart Campus Management], 4 January 2019, https://cloud.tencent.com/developer/article/1370766; 人民网 [People.com.cn], 'AI外教上英语课？小学课辅开启线上线下融合模式' ['AI Foreign Teachers in English Class? Online and Offline Integration Mode of Primary School Course Assistance Opens'], 1 June 2019, http://sc.people.com.cn/n2/2019/0621/c345167-33063596.html

142 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, no. 3, Spring 2019, pp. 7–8, https://thedisconnect.co/three/camera-above-the-classroom/. Hanwang Technology, now known as 'China's grandfather of facial recognition' began working on face recognition ID in part due to its early pivot away from fingerprint-scanning and towards contactless identity verification after the SARS epidemic. During the COVID-19 pandemic, the firm has claimed it can accurately use face recognition to identify people wearing masks. See: M. Pollard, 'Even Mask-Wearers Can Be ID'd, China Facial Recognition Firm Says', *Reuters*, 9 March 2020, https://www.reuters.com/article/us-health-coronavirus-facial-recognition/even-mask-wearers-can-be-idd-china-facial-recognition-firm-says-idUSKBN20W0WL; L. Lucas and E. Feng, 'Inside China's Surveillance State', *Financial Times*, 20 July 2018, https://www.ft.com/content/2182eebe-8a17-11e8-bf9e-8771d5404543

143 亿欧 [*Yi'ou*], '海风教育上线AI系统"好望角", 情绪识别是AI落地教育的重点方向?' ['Haifeng Education "Cape of Hope" AI System Goes Online, Is Emotion Recognition the Key Direction for Implementing AI in Education?'], 23 April 2018, https://www.iyiou.com/p/70791.html; *Sina*, '海风教育发布K12落地AI应用" 好望角" 借助情绪识别赋能教学' ['Haifeng Education Releases K-12 Implementation of AI Application "Cape of Good Hope", Empowers Teaching With Aid of Emotion Recognition'], 23 April 2018, http://edu.sina.com.cn/l/2018-04-23/doc-ifzfkmth7156702.shtml; Haifeng Education, '海风优势' ['Haifeng Advantage'], https://www.hfjy.com/hfAdvantage?uid=uid_1593910508_998072. Haifeng Education (海风教育) has a user base of over 4 million students, and focuses on K-12 online education and educational guidance in areas such as college preparedness. See: Haifeng Education, '海风教育好望角系统 打响'教育+AI'第一枪' ['Haifeng Education's Cape of Good Hope System Fires First Shot of "AI+Education"'], 13 February 2019, https://www.chengjimanfen.com/yiduiyifudao_xinwen/414

144 杭州网 [Hangzhou.com.cn], '智慧课堂行为管理系统上线 教室'慧眼'锁定你' ['Smart Classroom Behavior Management System Goes Online, Classroom's "Smart Eyes" Lock Onto You']; 17 May 2018, https://hznews.hangzhou.com.cn/kejiao/content/2018-05/17/content_7003432.html

145 新京报网 [*The Beijing News*], '杭州一中学课堂引入人脸识别"黑科技"' ['Hangzhou No. 11 Middle School Introduces "Black Technology" for Face Recognition'], 18 May 2018, http://www.bjnews.com.cn/news/2018/05/18/487458.html

146 *Sohu*, '智能错题本、人脸情绪识别、课堂即时交互、智慧云课堂 – – 联想智慧教育将为北京"新高"考赋' ['Smart Wrong Answer Book, Facial Emotion Recognition, Immediate Classroom Interaction, Smart Cloud Classroom – – Lenovo Smart Education Will Enable "New Gaokao" for Beijing'], 9 September 2019, https://www.sohu.com/a/339676891_363172; and 经济网 [*CE Weekly*], '联想打造全国首个科技公益教育平台 开播首课25万人观看' ['Lenovo Builds Country's First Science and Technology Public Welfare Education Platform, Broadcasts First Class to 250,000 Viewers'], 6 March 2020, http://www.ceweekly.cn/2020/0306/289041.shtml. A 错题本 (*cuòtíběn*) is a workbook containing incorrect answers to questions and explaining why they are wrong.

147 *Tianyancha* [天眼查], '蜜枣网：自主研发情绪智能分析系统，深度改变零售与幼教' ['Meezao: Independent Research and Development of Emotion Intelligence Analytical Systems, Deeply Changing Logistics and Preschool Education'], 28 June 2018, https://news.tianyancha.com/ll_i074g8rnrd.html

148 Meezao, 'AI技术将改进基础教育的方法与效率' ['AI Technology Will Change Methods and Efficiency of Basic Education'], 26 December 26 2019, http://www.meezao.com/news/shownews.php?id=62

149 网易 [NetEase], '新东方助力打造雅安首个AI双师课堂' ['New Oriental Helps Build Ya'an's First AI Dual Teacher Classroom'], 6 September 2018, https://edu.163.com/18/0906/11/DR14325T00297VGM.html

150 极客公园 [*GeekPark*], '「今天我的课堂专注度在三位同学中最高！」比邻东方「AI 班主任」用数据量化孩子课堂表现' ['"Today My Class Concentration Level is the Highest

Among Three Classmates!" Bling ABC New Oriental "AI Class Teacher" Uses Data to Quantify Children's Classroom Performance'], 2 November 2018, https://www.geekpark.net/news/234556

151 Taigusys Computing, '产品中心: AI课堂专注度分析系统' ['Product Center: AI Classroom Concentration Analysis System'], http://www.taigusys.com/procen/procen162.html

152 In Mandarin, TAL goes by the name 好未来 ('good future'). Originally named Xue'ersi (学而思), the company changed its name to TAL in 2013, but still retains the Xue'ersi name on products including the Magic Mirror. See: 天元数据 [*Tianyuan Data*], '揭秘中国市值最高教育巨头: 狂奔16年，靠什么跑出'好未来'' ['Unmasking China's Highest Market Value Education Giant: In a 16-Year Mad Dash, What to Rely On to Run Towards a "Good Future"?'], 10 June 2019, https://www.tdata.cn/int/content/index/id/viewpoint_102421.html

153 *Sohu*, '打造"未来智慧课堂"科技让教育更懂孩子 ['Create "Future Smart Classroom" Technology to Make Education Understand Children More'], 23 October 2017, https://www.sohu.com/a/199552733_114988

154 李保宏 [L. Baohong], '人工智能在中俄两国教育领域发展现状及趋势' ['The Status Quo and Trend of Artificial Intelligence in the Field of Education in China and Russia'], *Science Innovation*, vol. 7, no. 4, 2019, p. 134, http://sciencepg.org/journal/archive?journalid=180&issueid=180070

155 宁波新闻网 [*Ningbo News Network*], '威创集团发布儿童成长平台，宣布与百度进行AI技术合作' ['VTron Group Releases Child Growth Platform, Announces AI Technology Cooperation with Baidu'], 1 July 2020, http://www.nbyoho.com/news/1577803932239322942.html; 证券日报网 [*Securities Daily*], '威创CPO郭丹：威创对科技赋能幼教三个层次的认知与实践' ['VTron CPO Guo Dan: VTron's Three-Tiered Thinking and Practice In Empowering Preschool Education With Science and Technology'], 16 November 2018, http://www.zqrb.cn/gscy/gongsi/2018-11-16/A1542353382387.html

156 Taigusys Computing, '产品中心: AI课堂专注度分析系统' ['Product Center: AI Classroom Concentration Analysis System'], http://www.taigusys.com/procen/procen162.html

157 For announcement of Tsinghua collaboration, see: *PR Newswire*, 'TAL Lays out Future Education in Terms of Scientific & Technological Innovation at the Global Education Summit', 6 December 2017, https://www.prnewswire.com/news-releases/tal-lays-out-future-education-in-terms-of-scientific-technological-innovation-at-the-global-education-summit-300567747.html. For documentation of FaceThink's partnership with TAL, see: 芥末堆 [*Jiemodui*], 'FaceThink 获好未来千万级投资，将情绪识别功能引入双师课堂' ['FaceThink Receives Tens of Millions in Investments from TAL, Introduces Emotion Recognition Function In Dual-Teacher Classrooms']. 3 May 2017, https://www.jiemodui.com/N/70500; 铅笔道 [*Qianbidao*], '获好未来投资 30岁副教授AI识别20种表情 实时记录学生上课状态' ['Winning Investment From TAL, 30-Year-Old Associate Professor's AI Recognizes 20 Expressions and Records Students' In-Class Status in Real Time'], 9 May 2017, https://www.pencilnews.cn/p/13947.html

158 Meezao, '创新为本，AI为善 --- 蜜枣网发布幼儿安全成长智能系统' [Innovation-Oriented, AI for Good – Meezao Presents a Smart System for Children's Safe Growth'], 7 January 2019, http://www.meezao.com/news/shownews.php?id=36; 拓扑社 [Topological Society], '以零售场景切入，蜜枣网利用情绪识别分析用户喜好降低流失率' ['Entering from Retail Scenarios, Meezao Uses Emotion Recognition to Analyze User Preferences and Reduce Turnover Rate'], *QQ*, 15 May 2018, https://new.qq.com/omn/

159 For details of the RYB incident, see: C. Buckley, 'Beijing Kindergarten Is Accused of Abuse, and Internet Erupts in Fury', *The New York Times*, 25 November 2017, https://www.nytimes.com/2017/11/24/world/asia/beijing-kindergarten-abuse.html. Meezao, '创新为本，AI为善 --- 蜜枣网发布幼 儿安全成长智能系统' [Innovation-Oriented, AI for Good – Meezao Presents a Smart System for Children's Safe Growth'], 7 January 2019, http://www.meezao.com/news/shownews.php?id=36

160 Meezao, '蜜枣网创始人赵小蒙：不惑之年的‘大龄梦想者’' ['Meezao Founder Zhao Xiaomeng: "Old Dreamers" in their Forties'], 27 June 2018, http://www.meezao.com/news/shownews.php?id=34

161 证券日报网 [*Securities Daily*], '威创CPO郭丹：威创对科技赋能幼教三个层次的认知与实践' ['VTron CPO Guo Dan: VTron's Three-Tiered Thinking and Practice In Empowering Preschool Education With Science and Technology'], 16 November 2018, http://www.zqrb.cn/gscy/gongsi/2018-11-16/A1542353382387.html

162 Lenovo, '智能情绪识别' ['Intelligent Emotion Recognition'], https://cube.lenovo.com/chatdetail.html

163 *Sohu*, '智能错题本、人脸情绪识别、课堂即时交互、智慧云课堂 – – 联想智慧教育将为北京"新高考"赋' ['Smart Wrong Answer Book, Facial Emotion Recognition, Immediate Classroom Interaction, Smart Cloud Classroom – Lenovo Smart Education Will Enable "New Gaokao" for Beijing'], 9 September 2019, https://www.sohu.com/a/339676891_363172; 经济网 [*CE Weekly*], '联想打造全国首个科技公益教育平台 开播首课25万人观看' ['Lenovo Builds Country's First Science and Technology Public Welfare Education Platform, Broadcasts First Class to 250,000 Viewers'], 6 March 2020, http://www.ceweekly.cn/2020/0306/289041.shtml

164 网易 [*NetEase*], '新东方助力打造雅安首个AI双师课堂' ['New Oriental Helps Build Ya'an's First AI Dual Teacher Classroom'], 6 September 2018, https://edu.163.com/18/0906/11/DR14325T00297VGM.html. The playing up of New Oriental's expansion into rural markets is not new for the company; in 2015, it live-streamed classes from a prestigious Chengdu high school to rural areas, agitating rural teachers, who felt displaced, and stunning students, who realised how far behind urban schools their curriculum was. See: X. Wang, 'Buffet Life', *Blockchain Chicken Farm*, Farrar, Straus, and Giroux, 2020, p. 107. New Oriental is also known for its involvement in a college-admissions scandal; see: S. Stecklow and A. Harney, 'How top US Colleges Hooked Up with Controversial Chinese Companies', *Reuters*, 2 December 2016, https://www.reuters.com/investigates/special-report/college-charities/

165 Although this school's name is translated as 'middle school', other accounts indicate it serves students of high-school age'. *Sina*, 杭州一中学引进智慧课堂行为管理系统引热议' ['Hangzhou No. 11 Middle School Introduces Smart Classroom Behavior Management System'], 18 July 2018, http://edu.sina.com.cn/zxx/2018-07-18/doc-ihfnsvyz9043937.shtml. Hikvision is one of the companies the US government has sanctioned for its provision of surveillance equipment that is used to surveil China's Muslim ethnic minority in Xinjiang province.

166 *ThePaper* [澎湃], 葛熔金 [Ge Rongjin]. '杭州一高中教室装组合摄像头，分析学生课堂表情促教学改进' ['A High School in Hangzhou Equipped Classrooms with Combined Cameras, Analyzes Students' Facial Expressions in the Classroom to Improve Teaching'], 16 May 2018, https://www.thepaper.cn/newsDetail_forward_2133853. Photographs of the Smart Classroom Behavior Management System's user interface contain examples of other data the system analyses, e.g. 'Today's School-Wide Classroom Expression Data' ('今日全校课堂表情数据') and '全校课堂智能感知数据趋势图', 'Entire School Classroom Smart Sensing Data Trend Graph', and 'Analysis of Classroom Attention Deviation' (班级课堂专注度偏离分析). See: Hangzhou No. 11 Middle School, '未来已来！未来智慧校园长啥样？快来杭十一中看看' ['The Future is Here! What Will the Smart Campus of the Future Look Like? Come Quick and See at Hangzhou No. 11 Middle School'], WeChat, 9 May 2018, https://mp.weixin.qq.com/s/zvH3OZH3Me2QLQB5lPA3vQ

167 腾讯网 [*Tencent News*], "智慧校园"就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html

168 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, no. 3, Spring 2019, p. 10, https://thedisconnect.co/three/camera-above-the-classroom/ p. 10.

169 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, no. 3, Spring 2019, pp. 11–12, https://thedisconnect.co/three/camera-above-the-classroom/

170 *Ibid.*

171 腾讯网 [*Tencent News*], ' '智慧校园'就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html

172 *Sina*, '杭州一中学引进智慧课堂行为管理系统引热议' ['Hangzhou No. 11 Middle School Introduces Smart Classroom Behavior Management System'], 18 July 2018, http://edu.sina.com.cn/zxx/2018-07-18/doc-ihfnsvyz9043937.shtml

173 D. Lee, 'At This Chinese School, Big Brother Was Watching Students – and Charting Every Smile or Frown', *Los Angeles Times*, 30 June 2018, https://www.latimes.com/world/la-fg-china-face-surveillance-2018-story.html

174 *Ibid.*

175 腾讯网 [*Tencent News*], '"智慧校园" 就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html

176 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, no. 3, Spring 2019, p. 13, https://thedisconnect.co/three/camera-above-the-classroom/

177 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, No. 3, Spring 2019, p. 18, https://thedisconnect.co/three/camera-above-the-classroom/

178 腾讯网 [*Tencent News*], '"智慧校园" 就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html

179 *Ibid.*

180 *Ibid.*

181 Hangzhou No. 11 Middle School, '杭州第十一中"智慧课堂管理系统" 引争议 – – 课堂需要什么样的"高科技"' ['"Smart Classroom Management System" in Hangzhou No.11 Middle School Causes Controversy – What Kind of "High Tech" Does a Classroom Need?'], 8 June 2018, http://www.hsyz.cn/article/detail/idhsyz_6336.htm

182 *Ibid.*

183 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, no. 3, Spring 2019, p. 10, https://thedisconnect.co/three/camera-above-the-classroom/

184 网易 [NetEase], '新东方助力打造雅安首个AI双师课堂' ['New Oriental Helps Build Ya'an's First AI Dual Teacher Classroom'], 6 September 2018, https://edu.163.com/18/0906/11/DR14325T00297VGM.htm

185 岳丽丽 [Y. Lili], '好未来推出"WISROOM"智慧课堂解决方案，升级'魔镜' ['TAL Launches "WISROOM" Smart Classroom Solution, Upgrades "Magic Mirror"'], 猎云网 [*Lieyun Network*], 18 July 2018, https://www.lieyunwang.com/archives/445413

186 李保宏 [L. Baohong], '人工智能在中俄两国教育领域发展现状及趋势' ['The Status Quo and Trend of Artificial Intelligence in the Field of Education in China and Russia'], *Science Innovation*, vol. 7, no. 4, 2019, p.134, http://sciencepg.org/journal/archive?journalid=180&issueid=1800704; 张无荒 [Z. Wuhuang], '海风教育让在线教育进入智能学习时代 '好望角'AI系统发布 ['Haifeng Education Brings Online Education into the Era of Smart Learning, Releases "Cape of Good Hope" AI System'], *Techweb*, 23 April 2018, http://ai.techweb.com.cn/2018-04-23/2657964.shtml.

187 葛熔金 [G. Rongjin], '杭州一高中教室装组合摄像头，分析学生课堂表情促教学改进' ['A High School in Hangzhou Equipped Classrooms With Combined Cameras, Analyzes Students' Facial Expressions in the Classroom to Improve Teaching'], *ThePaper* [澎湃], 16 May 2018, https://www.thepaper.cn/newsDetail_forward_2133853

188 赵雨欣 [Z. Yuxin], '人工智能抓取孩子课堂情绪？在线教育还能这样玩' ['AI Can Capture Children's Emotions in the Classroom? Online Education Can Do This Too'], 成都商报 [*Chengdu Economic Daily*], 15 December 2017, https://www.cdsb.com/Public/cdsb_offical/2017-12-15/1629504657121871460404440244082080842222.html

189 Hangzhou No. 11 Middle School, '杭州第十一中"智慧课堂管理系统"引争议 – – 课堂需要什么样的"高科技"' ['"Smart Classroom Management System" in Hangzhou No.11 Middle School Causes Controversy – What Kind of "High Tech" Does a Classroom Need?'], 8 June 2018 https://www.cdsb.com/Public/cdsb_offical/2017-12-15/1629504657121871460404440244082080842222.html

190 *Ibid.*

191 腾讯网 [*Tencent News*], "智慧校园" 就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html

192 腾讯网 [*Tencent News*], '人在坐，AI在看' ['While People Sit, AI Watches'], 3 September 2019, https://new.qq.com/omn/20190903/20190903A09VG600.html

193 For instance, one article recounted a court case from 2018, which revealed that an employee of major AI company iFlytek illegally sold student data. The employee was in charge of a school-registration management system in Anhui province, and was reported to have sold data from 40,000 students. 腾讯网 [*Tencent News*], '人在坐，AI在看' ['While People Sit, AI Watches'], 3 September 2019, https://new.qq.com/omn/20190903/20190903A09VG600.html

194 '智慧课堂行为管理系统上线 教室"慧眼"锁定你' ['Smart Classroom Behavior Management System Goes Online, Classroom's "Smart Eyes" Lock Onto You']. 杭州网 [Hangzhou.com.cn]. May 17, 2018. https://hznews.hangzhou.com.cn/kejiao/content/2018-05/17/content_7003432.htm

195 腾讯网 [*Tencent News*], ' "智慧校园" 就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html; 新京报网 [*The Beijing News*], '杭州一中学课堂引入人脸识别'黑科技'' ['Hangzhou No. 11 Middle School Introduces "Black Technology" for Face Recognition'], 18 May 2018, http://www.bjnews.com.cn/news/2018/05/18/487458.html. The claim that the Smart Classroom Behavior Management System only displays data on groups rather than individuals is at odds with the description of the monitors teachers can see, which provide push notifications about which students are inattentive.

196 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, no. 3, Spring 2019, p. 10, https://thedisconnect.co/three/camera-above-the-classroom/

197 葛熔金 [G. Rongjin], '杭州一高中教室装组合摄像头，分析学生课堂表情促教学改进' ['A High School in Hangzhou Equipped Classrooms with Combined Cameras, Analyzes Students' Facial Expressions in the Classroom to Improve Teaching'], *ThePaper* [澎湃], 16 May 2018, https://www.thepaper.cn/newsDetail_forward_2133853

198 新京报网 [*The Beijing News*], '杭州一中学课堂引入人脸识别"黑科技"' ['Hangzhou No. 11 Middle School Introduces "Black Technology" for Face Recognition'], 18 May 2018, http://www.bjnews.com.cn/news/2018/05/18/487458.html. Hikvision's Education Industry director Yu Yuntao echoed Zhang's statement about the expression-recognition data only being used for teachers' reference; see: 腾讯网 [*Tencent News*], ' "智慧校园"就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html

199 L. Lucas and E. Feng, 'Inside China's Surveillance State', *Financial Times*, 20 July 2018, https://www.ft.com/content/2182eebe-8a17-11e8-bf9e-8771d5404543

200 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, no. 3, Spring 2019, p. 19, https://thedisconnect.co/three/camera-above-the-classroom/

201 *Sohu*, '打造"未来智慧课堂"科技让教育更懂孩子 ['Create "Future Smart Classroom" Technology to Make Education Understand Children More'], 23 October 2017, https://www.sohu.com/a/199552733_114988; Hangzhou No. 11 Middle School, '未来已来！未来智慧校园长啥样？快来杭十一中看看' ['The Future is Here! What Will the Smart Campus of the Future Look Like? Come Quick and See at Hangzhou No. 11 Middle School'], WeChat, 9 May 2018, https://mp.weixin.qq.com/s/zvH3OZH3Me2QLQB5lPA3vQ

202 腾讯网 [*Tencent News*], '"智慧校园"就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html. The same month that Hangzhou No. 11 launched its Smart Classroom Behavior Management System, the nearby Jiangsu Province Education Department, the Jiangsu Institute of Economic and Information Technology, and the Jiangsu Province Department of Finance co-released the *Guiding Opinions on Jiangsu Province Primary and High School Smart Campus Construction* (《江苏省中小学智慧校园建设指导意见(试行)》). This policy document specifically references smart classrooms as 'collecting teaching and learning behavior data throughout the entire process'. 江苏省教育厅 [Jiangsu Education Department], '关于印发智慧校园建设指导意见的通知' ['Notice on Printing and Distributing Guiding Opinions on Smart Campus Construction'], 23 May 2018, http://jyt.jiangsu.gov.cn/art/2018/5/23/art_61418_7647103.html

203 Taigusys Computing, '深圳安全监控系统进校园' ['Shenzhen Security Surveillance System Enters Campus'], 19 January 2019, http://www.taigusys.com/news/news120.html

204 新京报网 [*The Beijing News*], '杭州一中学课堂引入人脸识别"黑科技"' ['Hangzhou No. 11 Middle School Introduces "Black Technology" for Face Recognition'], 18 May 2018, http://www.bjnews.com.cn/news/2018/05/18/487458.html

205 新京报网 [*The Beijing News*], '杭州一中学课堂引入人脸识别"黑科技"' ['Hangzhou No. 11 Middle School Introduces "Black Technology" for Face Recognition'], 18 May 2018, http://www.bjnews.com.cn/news/2018/05/18/487458.html; Hangzhou No. 11 Middle School, '我校隆重举行"未来智慧校园的探索与实践"活动' ['Our School Held a Grand Activity of "Exploration and Practice of Future Smart Campus"'], 16 May 2018, http://www.hsyz.cn/article/detail/idhsyz_6308.htm. The 'smart cafeteria' food-monitoring project was undertaken by Zhejiang Primary and Secondary School Education and Logistics Management Association's Primary and Secondary School Branch, and the Hangzhou Agricultural and Sideline Products Logistics Network Technology Co. Ltd. See: 腾讯网 [*Tencent News*], '"智慧校园"就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html

206 *Sohu*, '打造"未来智慧课堂"科技让教育更懂孩子 ['Create "Future Smart Classroom" Technology to Make Education Understand Children More'], 23 October 2017, https://www.sohu.com/a/199552733_114988

207 36氪 [*36Kr*], '新东方发布教育新产品"AI班主任," 人工智能这把双刃剑，教育公司到底怎么用?' ['New Oriental Releases New Education Product "AI Teacher", A Double-Edged Sword of AI: How Can Education Companies Use It?'], 29 October 2018, https://36kr.com/p/1722934067201

208 Y. Xie, 'Camera Above the Classroom', *The Disconnect*, no. 3, Spring 2019, pp. 6, 21, https://thedisconnect.co/three/camera-above-the-classroom/

209 *Ibid.*

210 腾讯网 [*Tencent News*], '"智慧校园"就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html

3. Emotion Recognition and Human Rights

211 J. Ruggie, 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework', *Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and other Business Enterprises*, A/HRC/17/31, UN Human Rights Council, 17th Session, 21 March 2011, https://undocs.org/A/HRC/17/31

212 D. Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, A/HRC/38/35, UN Human Rights Council, 38th Session, 6 April 2018, para 10, https://undocs.org/A/HRC/38/35

213 A. Chaskalson, 'Dignity as a Constitutional Value: A South African Perspective', *American University International Law Review*, vol. 26, no. 5, 2011, p. 1382, https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=2030&context=auilr

214 C. McCrudden, 'Human Dignity and Judicial Interpretation of Human Rights', *The European Journal of International Law*, vol. 19, no. 4, 2008, http://ejil.org/pdfs/19/4/1658.pdf

215 ARTICLE 19, *Right to Online Anonymity: Policy Brief*, June 2015, https://www.article19.org/data/files/medialibrary/38006/Anonymity_and_encryption_report_A5_final-web.pdf. Also see: S. Chander, 'Recommendations for a Fundamental Rights-Based Artificial Intelligence Regulation', *European Digital Rights*, 4 June 2020, https://edri.org/wp-content/uploads/2020/06/AI_EDRiRecommendations.pdf

216 Article 17(1), ICCPR; Article 11, ACHR ('2. No one may be the object of arbitrary or abusive interference with his private life, his family, his home, or his correspondence […] 3. Everyone has the right to the protection of the law against such interference […]'). Also see: UN Human Rights Committee, General Comment No. 16 (Article 17, ICCPR), 8 April 1988, para 3, http://tbinternet.ohchr.org/Treaties/CCPR/Shared%20Documents/1_Global/INT_CCPR_GEC_6624_E.doc (noting that '[t]he term "unlawful" means that no interference can take place except in cases envisaged by the law', and that '[i]nterference authorised by States can only take place on the basis of law, which itself must comply with the provisions, aims and objectives of the Covenant'); *Necessary and Proportionate: International Principles on the Application of Human Rights to Communications Surveillance*, Principle 1, https://necessaryandproportionate.org/principles (these principles apply international human rights law to modern digital surveillance; an international coalition of civil society, privacy, and technology experts drafted them in 2013, and over 600 organisations around the world have endorsed them

217 UN High Commissioner for Human Rights, *The Right to Privacy in the Digital Age: Report of the United Nations High Commissioner for Human Rights*, A/HRC/39/29, UN Human Rights Council, 39th Session, 3 August 2018, https://undocs.org/A/HRC/39/29

218 UN Human Rights Council, *Report of the Special Rapporteur on the Right to Privacy, Joseph A. Cannataci*, A/HRC/34/60, 24 February 2017, para 17, https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/34/60

219 Article 12.3, International Covenant on Civil and Political Rights, 16 December 1996 (23 March 1976), https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx

220 ARTICLE 19, *The Global Principles on Protection of Freedom of Expression and Privacy*, 19 January 2018, http://article19.shorthand.com/

221 General Comment No. 34, CCPR/C/GC/3, para 10; J. Blocher, *Rights To and Not To*, *California Law Review*, vol. 100, no. 4, 2012, pp. 761–815 (p. 770).

222 UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc A/68/362. 4 September 2013.

223 UN Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights*, p. 15, https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

224 The right of peaceful assembly includes the right to hold meetings, sit-ins, strikes, rallies, events, or protests, both offline and online. See: UN High Commissioner for Human Rights, *Impact of New Technologies on the Promotion and Protection of Human Rights in the Context of Assemblies, Including Peaceful Protests: Report of the United Nations High Commissioner for Human Rights*, A/HRC/44/24, 24 June 2020, para 5, https://undocs.org/en/A/HRC/44/24

225 E. Selinger and A.F. Cahn, 'Did You Protest Recently? Your Face Might Be in a Database', *The Guardian*, 17 July 2020, https://www.theguardian.com/commentisfree/2020/jul/17/protest-black-lives-matter-database; *The Wire*, 'Delhi Police Is Now Using Facial Recognition Software to Screen "Habitual Protestors"', 29 December 2019, https://thewire.in/government/delhi-police-is-now-using-facial-recognition-software-to-screen-habitual-protestors

226 UN Human Rights Council, *Report of the Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association*, 17 May 2019, para 57, https://www.ohchr.org/EN/Issues/AssemblyAssociation/Pages/DigitalAge.aspx

227 Organization for Security and Co-operation in Europe, 'Right to be Presumed Innocent and Privilege against Self-Incrimination', *Legal Digest of International Fair Trials*, p. 99, https://www.osce.org/files/f/documents/1/f/94214.pdf#page=90

228 UN Human Rights Council, *Resolution on the Right to Privacy in the Digital Age*, A/HRC/34/L,7, 23 March 2017, page 3. https://digitallibrary.un.org/record/1307661?ln=en

229 凤凰网 [*iFeng News*], '打造保险科技转型急先锋 平安产险携多项AI技术亮相世界人工智能大会' ['Creating a Pioneer in the Transformation of Insurance Technology, Ping An Property Insurance Company Showcases Several AI Technologies at the World Artificial Intelligence Conference'], 29 August 2019, http://sn.ifeng.com/a/20190829/7693650_0.shtml, and Xinhua, "金融与科技加速融合迈入"智能金融时代"' ['Accelerate the Fusion of Finance and Technology Into the "Era of Smart Finance"'], 30 August 2019, http://www.xinhuanet.com/2019-08/30/c_1124942152.htm

230 *Tianyancha* [天眼查], '蜜枣网：自主研发情绪智能分析系统，深度改变零售与幼教' ['Meezao: Independent Research and Development of Emotion Intelligence Analytical Systems, Deeply Changing Logistics and Kindergarten Education'], 28 June, 2018, https://news.tianyancha.com/ll_i074g8rnrd.html, and Meezao official company website, "蜜枣网CEO赵小蒙：除了新零售，情绪智能识别还可以改变幼教" ["Meezao CEO Zhao Xiaomeng: In Addition to New Retail, Smart Emotional Recognition Can Also Change Preschool Education"], 19 July 2018, http://www.meezao.com/news/shownews.php?id=35

231 "杭州一中学课堂引入人脸识别‘黑科技’" ["Hangzhou No. 1 Middle School Introduces "Black Technology" for Face Recognition"]. 新京报网 [*The Beijing News*]. 18 May 2018, http://www.bjnews.com.cn/news/2018/05/18/487458.html

232 蔡村、陈正东、沈蓓蓓. [C. Cun, C. Zhengdong, and S. Beibei], '把握瞬间真实:海关旅检应用微表情心理学的构想' ['Grasp the Truth in an Instant: Application of Micro-Expressions Psychology in Customs Inspection of Passengers'],《海关与经贸研究》[*Journal of Customs and Trade*], no. 3, 2018, pp. 31, 33.

233 M. Beraja, D.Y. Yang, and N. Yuchtman, *Data-Intensive Innovation and the State: Evidence from AI Firms in China* (draft), 16 August 2020, http://davidyyang.com/pdfs/ai_draft.pdf

234 中国青年网 [Youth.cn], '"曲靖模式"先行项目 – – 翼开科技，成功在曲落地扎根 ['The First "Qujing Style" Project-EmoKit Technology Successfully Takes Root in Quluo'], 5 September 2019, http://finance.

youth.cn/finance_cyxfgsxw/201909/t20190905_12062221.htm. The article lays out a trajectory EmoKit plans to follow: the company would first have to raise 6 million yuan [USD 927,520] from angel investors, then pursue academic and market promotion activities with People's Public Security University of China and similar institutions, followed by approval to enter the Ministry of Public Security's procurement equipment directory (公安部装备采购名录), and would finally sell its products to South and Southeast Asian countries.

235 云涌 [*Yunyong (Ningbo News)*], '专访之三：看一眼就读懂你，甬企这双"鹰眼"安防科技够"黑"' ['Interview 3: It Can Understand You in One Glance, This Ningbo Company's Pair of "Hawk Eyes" Security Technology is "Black" Enough']. 4 May 2018, http://yy.cnnb.com.cn/system/2018/05/04/008748677.shtml

236 Hikvision, *Success Stories*, https://www.hikvision.com/content/dam/hikvision/en/brochures-download/success-stories/Success-Stories-2019.pdf. In case original source link is broken, please contact the authors for a copy.

237 *Infoweek*, 'Surveillance Networks Operated by China Spread Throughout Latin America', 7 August 2019, https://infoweek.biz/2019/08/07/seguridad-redes-de-vigilancia-china-latino-america/; *LaPolitica Online*, 'Huawei Lands in Mendoza to Sell its Supercameras with Facial Recognition and Big Data', 28 April 2018, https://www.lapoliticaonline.com/nota/112439-huawei-desembarca-en-mendoza-para-vender-sus-supercamaras-con-reconocimiento-facial-y-big-data/; T. Wilson and M. Murgia, 'Uganda Confirms Use of Huawei Facial Recognition Cameras', *The Financial Times*, 21 August 2019, https://www.ft.com/content/e20580de-c35f-11e9-a8e9-296ca66511c9; S. Woodhams, 'Huawei Says its Surveillance Tech Will Keep African Cities Safe but Activists Worry it'll Be Misused', *Quartz*, 20 March 2020, https://qz.com/africa/1822312/huaweis-surveillance-tech-in-africa-worries-activists/; B. Jardine, 'China's Surveillance State Has Eyes on Central Asia', *Foreign Policy*, 15 November 2019, https://foreignpolicy.com/2019/11/15/huawei-xinjiang-kazakhstan-uzbekistan-china-surveillance-state-eyes-central-asia/

238 Original source link is broken, please contact authors for a copy.

239 UN Human Rights Council, *Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis*, 18 June 2020, https://www.ohchr.org/EN/Issues/Racism/SRRacism/Pages/SRRacismThematicReports.aspx

240 For examples of exceptions, see: 腾讯网 [*Tencent News*], '"智慧校园"就这么开始了，它是个生意，还是个问题?' ['Is This Kind of Start to "Smart Campuses" a Business or a Problem?'], 30 May 2018, https://new.qq.com/omn/20180530/20180530A03695.html, 马爱平 [M. Aiping], 'AI不仅能认脸，还能"读心"' ['AI Doesn't Just Read Faces, It Can Also "Read Hearts"'], *Xinhua*, 17 June 2020, http://www.xinhuanet.com/fortune/2020-06/17/c_1126123641.htm

241 《南京日报》 [*Nanjing Daily*], '多个城市已利用AI读心加强反恐安防' ['Several Cities Have Used AI Mind Reading to Strengthen Counterterrorist Security'], 29 September 2018, http://njrb.njdaily.cn/njrb/html/2018-09/29/content_514652.htm.

242 L. Rhue, 'Racial Influence on Automated Perceptions of Emotions', SSRN, November 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765; L. Rhue, 'Emotion-reading Tech Fails the Racial Bias Test', 3 January 2019, *The Conversation*, https://theconversation.com/emotion-reading-tech-fails-the-racial-bias-test-108404. Megvii is one of the companies sanctioned in the US for supplying authorities in Xinjiang province with face recognition cameras used to monitor Uighur citizens.

243 段蓓玲 [D. Beiling], '视频侦查主动预警系统应用研究' ['Applied Research on Active Early Warning System for Video Investigations'], 《法制博览》 [*Legality Vision*], no. 16, 2019 p. 65.

244 For Baidu's API, see: "人脸检测" [Face Detection"]. Baidu official company website.

16 April 2020. http://ai.baidu.com/ai-doc/FACE/yk37c1u4t. For the Taigusys Computing API, see: "表情识别对外开放接口 v1.0" ["Facial Expression Recognition Open Interface v1.0"]. Taigusys Computing official company website, 16 December 2019. http://www.taigusys.com/download/download105.html

245 R. el Kaliouby, 'Q&A with Affectiva: Answers to Your Most Emotional Questions', *Affectiva* [blog], 3 March 2017, https://blog.affectiva.com/qa-with-affectiva-answers-to-your-most-emotional-questions

246 For Taigusys, see: '产品中心：智慧监所情绪识别预警分析系统' ['Product Center: Smart Prison Emotion Recognition Early Warning Analysis System'], Taigusys Computing official company website, http://www.taigusys.com/procen/procen160.html. For EmoKit, see: "他用AI情感算法来做"测谎仪"已为网贷公司提供反骗贷服务 获订单300万" ["He Used AI Emotion Algorithms to Make a 'Lie detector,' Has Already Provided Anti-fraud Services to Online Lending Companies and Won 3 Million Orders"]. 铅笔道 [Qianbidao], 23 July 2018. https://www.pencilnews.cn/p/20170.html

247 Meezao, '蜜枣网创始人赵小蒙：不惑之年的"大龄梦想者"' ['Meezao Founder Zhao Xiaomeng: "Older Dreamers" in Their Forties'], 27 June 2018, http://www.meezao.com/news/shownews.php?id=34

248 M. Whittaker, *et al*., *Disability, Bias, and AI*, AI Now Institute, November 2019, p.13, https://ainowinstitute.org/disabilitybiasai-2019.pdf

249 L. Yue, Z. Chunhong, T. Chujie, Z. Xiaomeng, Z. Ruizhi, and J. Yang, 'Application of Data Mining for Young Children Education Using Emotion Information', *DSIT '18: Proceedings of the 2018 International Conference on Data Science and Information Technology*, July 2018, p. 2, https://doi.org/10.1145/3239283.3239321. At the time of the paper's publication, aside from Meezao CEO Zhao Xiaomeng, all the paper's authors were affiliated with Beijing University of Posts and Telecommunications, which has been cited, alongside Tsinghua University and Capital Normal University, as a collaborator on Meezao's technology for measuring students' concentration. See: Meezao, '蜜枣网联合微软推出全球首例人工智能儿童专注能力分析系统' ['Meezao Joins Microsoft in Launch of World's First AI Child Concentration Analysis System'], 8 January 2019, http://www.meezao.com/news/shownews.php?id=45

250 L. Yue, Z. Chunhong, T. Chujie, Z. Xiaomeng, Z. Ruizhi, and J. Yang, 'Application of Data Mining for Young Children Education Using Emotion Information', DSIT '18: *Proceedings of the 2018 International Conference on Data Science and Information Technology*, July 2018, p. 2, https://doi.org/10.1145/3239283.3239321

251 *Ibid.*

252 C. Bergstrom and J. West, 'Case Study: Criminal Machine Learning', *Calling Bullshit*, University of Washington, https://www.callingbullshit.org/case_studies/case_study_criminal_machine_learning.html. A 2020 paper from Harrisburg University similarly purporting to predict criminality from faces was barred from publication after over 1,000 researchers signed a petition letter; see: S. Fussell, 'An Algorithm That "Predicts" Criminality Based on a Face Sparks a Furor', *WIRED*, 24 June 2020, https://www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/

253 中关村国家自主创新示范区 [Zhongguancun National Independent Innovation Demonstration Zone], 'EmoAsk AI多模态智能审讯辅助系统' ['EmoAsk AI Multimodal Smart Interrogation Auxiliary System'], http://www.zgcnewth.com/qiyefuwu/productInfo/detail?eid=1&pid=11425

254 彭玉伟 [P. Yuwei], '理性看待微表情分析技术在侦讯工作中的应用' ['Rationally Treat the Application of Micro-Expressions Analysis Technique in Interrogation'], 《湖南警察学院学报》[*Journal of Hunan Police Academy*], vol. 26, no. 2, April 2014, pp. 12–13.

255 南通安全防范协会 [Nantong Security and Protection Association], '从情绪识别谈校园安防的升级' ['Discussing Improvement of Campus Safety With Emotion Recognition'], 4 May 2018, http://www.jsntspa.com/contents/68/363.html

256 王鹏, 马红平 [W. Peng and M. Hongping], '公共场所视频监控预警系统的应用' ['Research on Application of Video Monitoring and Warning System in Public Places'],《广西警察学院学报》[*Journal of Guangxi Police College*], vol. 31, no. 2, 2018, pp. 42–45.

4. China's Legal Framework and Human Rights

257 For more context on this, see: UN Human Rights Council, *Report of the Working Group on the Universal Periodic Review: China*, A/HRC/11/255, October 2009. Also see: W. Zeldin, 'China: Legal Scholars Call for Ratification of ICCPR', *Global Legal Monitor*, 2 February 2008, https://www.loc.gov/law/foreign-news/article/china-legal-scholars-call-for-ratification-of-iccpr/; V. Yu, 'Petition Urges NPC to Ratify Human Rights Treaty in China', *South China Morning Post*, 28 February 2013, https://www.scmp.com/news/china/article/1160622/petition-urges-npc-ratify-human-rights-treaty-china; *Rights Defender*, 'Nearly a Hundred Shanghai Residents Called on the National People's Congress to Ratify the International Covenant on Civil and Political Rights (Photo)', 23 July 2014, http://wqw2010.blogspot.com/2013/07/blog-post_8410.html

258 Information Office of the State Council, The People's Republic of China, *National Human Rights Action Plan for China, 2016–2020*, 1st ed., August 2016, http://www.chinahumanrights.org/html/2016/POLITICS_0929/5844.html

259 For tracing how this intention has spanned multiple reports, see the 2012–2015 plan: Permanent Mission of the People's Republic of China to the United Nations Office at Geneva and Other International Organizations in Switzerland, *National Human Rights Action Plan for China (2012–2015)*, 11 June 2012, http://www.china-un.ch/eng/rqrd/jblc/t953936.htm#:~:text=The%20period%202012%2D2015%20is,it%20is%20also%20an%20important

260 Article 40 of the Chinese Constitution. The freedom and privacy of correspondence of citizens of the People's Republic of China are protected by law. No organisation or individual may, on any ground, infringe upon the freedom and privacy of citizens' correspondence – except in cases where, to meet the needs of state security or of investigation into criminal offences, public security or procuratorial organs are permitted to censor correspondence in accordance with procedures prescribed by law.

261 Q. Zhang, 'A Constitution Without Constitutionalism? The Paths of Constitutional Development in China',. *International Journal of Constitutional Law*, vol. 8, no. 4, October 2010, pp. 950–976, https://doi.org/10.1093/icon/mor003, https://academic.oup.com/icon/article/8/4/950/667092

262 J. Ding, 'ChinAI #77: A Strong Argument Against Facial Recognition in the Beijing Subway', *ChinAI*, 10 December 2019, https://chinai.substack.com/p/chinai-77-a-strong-argument-against

263 E. Pernot-Leplay, 'China's Approach on Data Privacy Law: A Third Way Between the US and EU?', *Penn State Journal of Law and International Affairs*, vol. 8, no. 1, May 2020, https://elibrary.law.psu.edu/cgi/viewcontent.cgi?article=1244&context=jlia. Also see: Y. Wu *et al*., 'A Comparative Study of Online Privacy Regulations in the U.S. and China', *Telecommunications Policy*, no. 35, 2011, pp. 603, 613.

264 P. Triolo, S. Sacks, G. Webster, and R. Creemers, 'China's Cybersecurity Law One Year On', *DigiChina*, 30 November 2017, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinas-cybersecurity-law-one-year/

265 R. Creemers, P. Triolo, and G. Webster, 'Translation: Cybersecurity Law of the People's Republic of China (Effective 1 June 2017)', *DigiChina*, 29 June 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/

266 G. Greenleaf and S. Livingston, 'China's New Cybersecurity Law – Also a Data Privacy Law?', *UNSW Law Research Paper*, no. 17–19; *144 Privacy Laws & Business International Report*, no. 1–7, 1 December 2016, https://papers.ssrn.

com/sol3/papers.cfm?abstract_id=2958658

267 *Phys Org*, 'China's Alibaba Under Fire Over Use of Customer Data', 5 January 2018, https://phys.org/news/2018-01-china-alibaba-customer.html

268 S. Sacks, Q. Chen, and G. Webster, 'Five Important Takeaways from China's Draft Data Security Law', *DigiChina*, 9 July 2020, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/five-important-take-aways-chinas-draft-data-security-law/

269 R. Creemers, M. Shi, L. Dudley, and G. Webster, 'China's Draft "Personal Information Protection Law" (Full Translation)', *DigiChina*, 21 October 2020, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinas-draft-personal-information-protection-law-full-translation/; G. Webster and R. Creemers, 'A Chinese Scholar Outlines Stakes for New "Personal Information" and "Data Security" Laws (Translation)', *DigiChina*, 28 May 2020, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinese-scholar-outlines-stakes-new-personal-information-and-data-security-laws-translation/

270 P. Triolo, S. Sacks, G. Webster, and R. Creemers, 'China's Cybersecurity Law One Year On', *DigiChina*, 30 November 2017, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinas-cybersecurity-law-one-year/

271 For an official translation of the 2020 version, see: *State Administration for Market Supervision of the People's Republic of China and Standardization Administration of the People's Republic of China, Information Security Technology* – Personal Information (PI) Security Specification: National Standard for the People's Republic of China, GB/T 35273–2020, 6 March 2020, https://www.tc260.org.cn/front/postDetail.html?id=20200918200432

272 S. Sacks, 'New China Data Privacy Standard Looks More Far-Reaching than GDPR', *CSIS*, 29 January 2018, https://www.csis.org/analysis/new-china-data-privacy-standard-looks-more-far-reaching-gdpr

273 S. Sacks, M. Shi, and G. Webster, 'The Evolution of China's Data Governance Regime', *DigiChina*, 8 February 2019, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/china-data-governance-regime-timeline/

274 J. Ding and P. Triolo, 'Translation: Excerpts from China's "White Paper on Artificial Intelligence Standardization"', *DigiChina*, 20 June 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-excerpts-chinas-white-paper-artificial-intelligence-standardization/

275 J. Ding, 'ChinAI #84: Biometric Recognition White Paper 2019', *ChinAI*, 2 March 2019, https://chinai.substack.com/p/chinai-84-biometric-recognition-white

276 G. Webster, R. Creemers, P. Triolo, and E. Kania, 'Full Translation: China's "New Generation Artificial Intelligence Development Plan" (2017)', *DigiChina*, 1 August 2017, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/

277 M. Shi, 'Translation: Principles and Criteria from China's Draft Privacy Impact Assessment Guide', *DigiChina*, 13 September 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-principles-and-criteria-from-chinas-draft-privacy-impact-assessment-guide/

278 G. Webster, 'Translation: Chinese AI Alliance Drafts Self-Discipline "Joint Pledge"', *DigiChina*, 17 June 2019, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/

279 L. Laskai and G. Webster, 'Translation: Chinese Expert Group Offers "Governance Principles" for "Responsible AI"', *DigiChina*, 17 June 2019, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/

280 BAAI, *Beijing AI Principles*, 28 May 2019, https://www.baai.ac.cn/news/beijing-ai-principles-en.html

ARTICLE 19
Free Word Centre
60 Farringdon Road
London EC1R 3GA
United Kingdom

article19.org

Goradia SG, J Alzheimers Neurodegener Dis 2019, 5: 023
DOI: 10.24966/AND-9608/100023

HSOA Journal of Alzheimer's and Neurodegenerative Diseases

Review Article

The Quantum Theory of Entanglement and Alzheimer's

Shantilal Gangadas Goradia*

President, Gravity Research Institute, Inc.,Shantiniketan 1, 983 David Walker Dr., Tavares, FL, USA

Abstract

A unifiable quantum theory of gravity should link to information and include biology as well as entanglement. There are many quantum theories of gravity linked to Einstein's theory of general relativity. We link to Newtonian gravity to show new horizons of information, entanglement, coherence, synchrony, consciousness, Alzheimer's, strong coupling, dark matter, geological experiences, Brownian motion, and more in our referred papers. A particle has been observed to exist at more than one place at the same time. We remember Stephen Hawking's view in our words that the characteristic of a true theory is that it looks true from different angles, we say, like the roundness of the earth. The world wants to focus her attempts to cure the horrible decease of Alzheimer's, while its cause is unknown. Understanding the cause may help to find the cure. The important horizons are consequences of our quantum theory of gravity. We cannot help referring to it. So we do.

Keywords: Non-locality; Alzheimer's; Consciousness; Quantum gravity; Entanglement

Introduction

The reasons how gravity links to strong force, why it is weak, is probabilistic, incorporates information, explains dark matter when quantized, relates to consciousness, connects to the ancient views, and entangles, leading to coherence in the brain are progressively documented in our open access articles [1-8] with related background information. Here, we support our probabilistic (quantum) theory of gravity, with the views of recognized dignitaries, provide some hysterical perspectives and respond to the issue in a popular journal about the lack of mathematics to support an observation that Nature sends information about a forthcoming earthquake to the sky [9], that we cannot pick up ahead of its occurrence. We add our view on the potential cause of Alzheimer's.

***Corresponding author:** Shantilal Gangadas Goradia, Gravity Research Institute, Inc., Shantiniketan 1, 983 David Walker Dr., Tavares, FL, USA, Tel: +1 5748556113; E-mail: sg@gravityresearchinstitute.org

Citation: Goradia SG (2019) The Quantum Theory of Entanglement and Alzheimer's. J Alzheimers Neurodegener Dis 5: 023.

Received: July 8, 2019; **Accepted:** July 18, 2019; **Published:** July 25, 2019

Entanglement

Quantum entanglement is defined as a physical phenomenon in which the measurement done on one elementary particle will instantaneously affect the state of another elementary particle located far away unrestricted by the speed of light limitation in Einstein's theory of special relativity. In reality, we say, there are not only two particles, but multiple particles that can show the phenomenon of entanglement. According to our quantum theory of gravity, one particle can interact with two or more (multiple) particles instantaneously on a probabilistic basis. Last year we published our view about "The Quantum Theory of Entanglement and Brain Physics"[8]. There, we used the phrase, "multiple particles can reciprocate", which involves the author's concept of the reality of Nature relating to elementary particles. The reciprocation in our concept is instantaneous leading to the fundamental cause of entanglement. The multiple particles, we implied in [5] on dark matter, are separate space times. Having said that, we are delighted to meet legendary figure, Feynman's 1957 view about superposition (entanglement), marvelously dug out and brought forth to all of us by Tim Folger of Scientific American [10].

Feynman's View

"Feynman argued that if gravity is indeed a quantum phenomenon, a superposition of a particle in two places at once would create two separate gravitational fields in case of a small mass in a quantum superposition, two different spacetimes would coexist side by side, almost like two separate universes, a state of affairs that should not exist in Einstein's theory [10]". Our article on dark matter in [5] implies that each quantum particle is entangled with multiple quantum particles in the universe and they all create their own gravitational forces. The entangled particles are the superposition. Feynman argued that only quantum phenomena can be entangled [10]. Therefore we meet Feynman's argument. We explained why we quantize Newtonian gravity in [8].

Elaboration

Since "The Inverse Square" is a common denominator of our quantum theory of gravity and that of the Newtonian gravity at short scales, we say our inverse square makes the classical one come out that way in addition to its ability to show entanglement on a logistical basis, and derive strong coupling to supplement our older derivations in [1-3]. Newton was in doubt about the duplicate reciprocal (inverse square law) [3]. We respond to Newton' doubt on his inverse square law by setting his inverse square on a quantum mechanical Planck scale, probabilistic basis, generating an equivalent of a homogenized mixture of the two non-unifying (quantum and classical) theories. Newton could not have known quantum reality, dark matter, entanglement and ever increasing knowledge of quantum weirdness.

A negative (dislikable) point is that ours has a very weird implication that each particle must be like a centipede with invisible legs and multiple existences of its particle like legs-tips and that the centipede extends the legs all over the universe instantly to interact with

other particles as if she uses the legs for a swing dance with other particles. We stick with the positive aspect of that dislikable point since it yields the overdue explanation of entanglement which must play its part also in the brain. Without Einstein's help, Heisenberg's paper on uncertainty could not be published. We can see the reason why: In a simplified illustration it leads to a weird conclusion that the elementary particles are behaving like worms.

Our theory of quantum gravity claims that gravity is the cumulative effect of the long range manifestations of the constants of Nature playing their part to create the coherence in the brain [8]. The explanation extends to synchronicity. Since the difference between synchronicity and coherence is that of scale per book [11] p 220. The book adds "If dark matter and dark energy have genuine physical properties, associated dark information must also exist [11] p 323. "True, it does exist per the subsection "Information Paradox" of [5] on dark matter".

While talking about two nodes in the brain, Dr. Nunez writes in his (2016) book [11] p 215: "..the label "functionally connected" assumes nothing about the cause or causes of this statistical relationship." We have now provided the cause of a statistical relationship implicitly and briefly in 2018 [8] in subsection "Global Information". The statistical relationship is analogous to the numerical language of Nature in our article [6].

The ON and OFF particle interactions every Planck time are like the binary system of information per our quantum theory of gravity; they imply that the universe speaks in integer numbers of Planck units. The weirdness of quantum mechanics to a common man is why not half a Planck unit? Explanation of money no less than a penny, makes the questioner swallow the answer.

Geology and Information

10/2018 Issue of Scientific American article [9] shows lack of a physical basis for the earth quake information reaching the sky prior to the earth quake. Here, we try to substantiate our abstract [12] about the subject information. The drastic relative movements of (1) the high order of magnitude of the particles in huge subcontinent sized tectonic plates, below the surface of the earth prior to the earthquake would change their individual distances from (2) the air particles floating above the earth, creating the changes in the ON and OFF interactions between (1) and (2) per our quantum theory of gravity, resulting in the drastic redistribution of the forces, expressible in Planck scale, of the constants of Nature in the air that birds must be capable of sensing and translating the implicit information into the possibility of a forth coming disaster. Entanglement of particles would create their own gravity effects, consistent with Feynman's View above. If technology can pick up the subject anomaly, it may be able to forecast earthquakes.

We realize it is a big "If", since the forces from the constants of Nature like the strong coupling become so diluted that they become practically indistinguishable at short distances of 1000 fm (about the radius of an atom). Our visualization of the dilution effect, linking strong force to weak gravity was at the base of our question to Dr. Weinberg in the following paragraph.

Consciousness/Alzheimer's

Nobel Laureate Dr. Weinberg has an interesting chapter in his book titled, "The Trouble with Quantum Mechanics" [13], citing Nobel Laureate Eugene Wagner's view that it was impossible to formulate the laws of quantum mechanics consistently without referring to the consciousness. We implicitly put consciousness on a scientific basis with our "principle of reciprocity" in the phrase: "multiple particles can reciprocate" in [8]. We welcomed Wagner's view, since it gave us a platform to stand on before writing the book [14], prepared primarily for a hand out at our 2011 oral presentation in Stockholm for a conference on consciousness. Here, we again thank Dr. Weinberg with due respect for a humorously encouraging, well received by thousands, and congratulating answer to our question related to our crave for unifying strong coupling with gravitation, at the APS Centennial Meeting, Atlanta, Georgia in March 1999. Our key point, now, in [8] is that gravity is fundamentally linked to nature's information system enhancing coherence in the brain and synchrony noticed elsewhere. We cannot rule out the possibility that Einstein and Satyendra Bose of India got the idea of bosons (which led to the discovery of Higgs Boson, the God particle) at the same instant. Entanglement must exist everywhere in the universe; we see no reason why the brain should be an exception. Amyloid Plaques and tau tangles are considered the cause of Alzheimer's [15]. If so, the search for the cure of Alzheimer's may reduce to finding an answer to the question: "how to nullify the information created by such proteins and plaques".

According to our quantum theory of gravity, one particle can interact with multiple particles in the universe instantaneously across the universe on a probabilistic basis. The probabilities of interaction of two particles, D1 and D2, non-zero, Planck Length distances apart , are squares of 1/D1 and 1/D2, such that, the combined probability would be the square of 1/(D1 x D2), which would be man's version of probabilistic mathematics in general, devoid of consciousness. If we combine the result with what quantum theory lacks per the view of Eugene Wagner, we have to think of the conscious mind of the particle involved. If the interacting particles stay engaged, the low probability event in general would become a longer term realistic event called entanglement, consistent with the implicit principle of reciprocation in [8]. The implicit body mind link at quantum scale, we say, manifests to 7/2019 issue of the Scientific American [16] high- lighted upfront as "How the Mind Arises-Network interactions in the brain create thought". The particle interactions must create gravity waves per equation 1 of [5], if not we would not have gravity. According to our quantum theory of gravity, the particle interactions are ON and OFF every Planck time and they would generate gravity waves with information, consistent with our information paradox in [8], making it needless to resort to the controversy of the conventional black hole information paradox in physics.

The disruption created by the age related plaques and protein deposits (structural changes) would add noise to the otherwise normal distribution of information in the orchestral music inside the brain, supplemented by age related vascular changes analogous to the structural changes. Since the probability of interaction in our quantum theory of gravity goes down with increasing separation of particles, the closer modules are liable to be more communicative, unless the consciousness aspect intervenes drastically in the network of brain communications. Some physicists including late Nobel Laureate John Bell believed that entanglement violates the spirit of the relativity theory [17]. Just because experimentalists observed two (or three) particles entangled at any one instant, the theorists cannot preclude the

possibility that multiple particles, say 20 out of 10E80 particles in the universe, can be entangled at an instant per the spirit of our theory, accounting for the dark matter in [5]. Since these 20 entangled particles belong to one system, there should not be even apparent violation of the uncertainty principle, as EPR had suggested according to the argument of Dr. Yanhua Shih in [17].

The reputed journal Nature did not test the brain of slaughtered pigs for consciousness [18]. Their amazing job succeeding to maintain the brain structure resulted in maintaining the brain's vitality. The probabilistic nature of our quantum gravity and related information exchanges as a function of the separations between particles must maintain the vitality so long as the brain structure is maintained as it did. The cat's implicit prediction of the forth coming demise of a nursing home resident as practiced in America is obviously not based on the cat looking at the brains of the residents, requiring a multidisciplinary study of probabilistic realities such as this.

We are delighted to see the pre-thinking of Feynman expressed in his famous quote about the mystical number (137) interpreted as potentially related to some natural logarithm matching our derivation of 137 in [4]. Regardless, we say that our derivation is not just a coincidence; it supports spin based information of the universe. The natural logarithm of probabilities we used there is not only entropy per Boltzmann, but also information per Shannon, the Guru of information theory. The number 137 is about electron/photon interaction called fine structure constant involved in the generation of energy from food we eat. We feel it is somewhat noteworthy that age related structural changes in the brain could adversely affect the locations where electron interactions play their part if it impacts Alzheimer's.

Meeting of Minds

Considering age reflects wisdom, we look at later views of Einstein. Regarding his EPR argument against the completeness of quantum theory, Einstein confessed to Schrodinger that the paper was written by Podolsky. It did not come out in the end so well [19]. Einstein was, in his final years, a realist, not a determinist [20]. We cannot answer questions about the fundamental cause of thought, anger etc. reminding us of popular books like [21,22] and Scientific American article [16].

Conclusion

An elementary particle has multiple existences each interacting with multiple existences of other elementary particles creating entanglement in the universe and coherence in the brain. The quantum theory is incomplete. A true quantum theory of gravity must show how it can fill the gap and show overall unification. We show ours as most promising.

References

1. Goradia S (2004) Gravity and strong force: Potentially linked by quantum wormholes. Indian Journal of Theoretical Physics 52: 143-149.
2. Goradia S (2006) Why is gravity so weak? J of Nuclear Radiation Physics 1: 107-118.
3. Goradia S (2012) Newtonian gravity in natural units. Journal of Physical Science and Application 2: 265-268.
4. Goradia S (2015) Decoding the information of life. J of Physical Science and Application 5: 191-195.
5. Goradia S (2015) Dark matter from our probabilistic gravity. J of Physical Science and Application 5: 373-376.
6. Goradia S (2016) Quantum consciousness-The road to reality. Journal of Life Sciences 10: 1-4.
7. Goradia S (2017) The emperor's mind in a nut shell. J of Life Sciences 11.
8. Goradia S (2018) The quantum theory of entanglement and brain physics. Journal of Clinical Review&Case Reports 3: 7.
9. Erik V (2018) Earthquakes in the sky. Scientific American.
10. Folger T (2019) Quantum gravity in the lab. Scientific American.
11. Nunez PL (2016) The new science of consciousness-Exploring the complexity of brain, mind, and self. (Prometheus Books, 59, John Glenn Drive, Amherst, New York).
12. Goradia S (2019) American physical society, Abstract: APR19-000103, Quantum Enigmas: Physics Encounters Consciousness & Earthquake Hints.
13. Weinberg S (2018) Third thoughts. The Belknap Press of Harvard University Press, Cambridge, Massachusetts.
14. Goradia S (2011) Quantum consciousness-The road to reality (Bloomington, IN, Author House).
15. Sherzai D, Sherzai A (2017) The Alzheimer's solution, (HarperCollins Publishers, 195 Broadway, New York, NY 10007).
16. Bassett DS, Bertolero M (2019) How Matter becomes Mind-The new discipline of network neuroscience yields a picture of how mental activity arises from carefully orchestrated interactions among different brain areas, illustrated by Mark Ross Studio, Scientific American.
17. Aczel AD (2002) Entanglement-The Greatest Mystery in Physics, (Four Walls, Eight Windows, New York, NY 100100) 252.
18. Vrselja Z, Daniele SG, Silbereis J, Talpo F, Morozov YM, et al. (2019) Restoration of brain circulation and cellular functions hours post- mortem, Nature 568: 336-343.
19. Howard D (1985) Studies in the history and Philosophy of Science Part A 16: 171-201.
20. Del Santo F (2019) Starving for realism, not for determination: Historic misconceptions on Einstein and Bohm. A Publication of the American Physical Society 28: 5.
21. Alexander E, Newell K (2017) Living in a mindful universe-A Neurosurgeon's Journey into the Heart of Consciousness, (RODALE).
22. Church D (2018) Mind to matter-The astonishing science of how your brain creates material reality, (HAY HOUSE, INC.).

World Journal of Condensed Matter Physics, 2015, 5, 244-248
Published Online August 2015 in SciRes. http://www.scirp.org/journal/wjcmp
http://dx.doi.org/10.4236/wjcmp.2015.53025

Entanglement: A Modern Aspect of Nature

Jens Cordelair

Birkenweg 2, Elmenhorst, Germany
Email: Jens.Cordelair@freenet.de

Received 21 July 2015; accepted 22 August 2015; published 25 August 2015

Abstract

The intention of this paper is to provide an easy to understand introduction to the peculiarities of entangled systems. A novel description for strong (mass entanglement) and weak (spin-or-bital and thermal entanglement) quantum entangled particles is discussed and applied to the phenomena of superconductivity, superfluidity and ultracold gases. A brief statement about how to represent the physical reality of quantum-entanglement as Quantum-Field-Theory (QFT) is noted.

Keywords

Quantum Entanglement, Superconductivity, Wave Particle Duality, Cooper-Pairs, Superfluidity, Ultracold Gases

1. Introduction

The word "nature" is derived from the Latin word natura with the physical meaning of "essential quality" or "innate disposition". In this sense I would like to show you how conservation laws and entanglement are inevitable parts of our physical thoughts.

A conservation law states that a particular measurable property of a physical system doesn't change as the system evolves, where entanglement describes the correlated evolution of the whole physical system to retain these conservation laws. In classical physics conservation of energy, momentum, angular momentum, mass and electric charge are common conservation laws. In particle physics other conservation laws such as baryon number, lepton number and strangeness apply to properties of subatomic particles that are invariant during an interaction.

In the following I want to introduce a novel description for strong (mass entanglement) and weak (spin-orbital and thermal entanglement) quantum entangled particles and to present some applications for the concept of quantum entanglement. In case of strong entangled particles the entanglement can't be shared with its environment, while weak entangled particles such as cooper-pairs or Bose-Einstein-condensates (BEC) can easily change its shape, where only the overall entanglement stays the same.

2. Particle Definition on the Basis of Entanglement

2.1. Mass Entanglement and Wave-Particle Duality

A fundamental aspect of physical thoughts is the principle of a homogeneous time development. As a result its influence on the interpretation of natural phenomena is very imperative. Formally the principle of a homogeneous time will be represented by the law of conservation of energy. Einstein showed in his theory of special relativity [1], that the mass m of a body is equivalent to the energy E in a proper measure, where c is the speed of light

$$E = mc^2 . \tag{1}$$

This means one can convert mass into energy (nuclear fission in the sun) and energy into mass (particle generation in high energy physics).

Another aspect of matter and energy was postulated by Louis de Broglie [2] in 1924 where he stated that matter should behave like light waves, also featuring interference. The simplest type of these matter waves is a plain and monochrome wave, where the energy- and momentum-distribution is restricted to a single value with ħ the Plank constant, P the momentum, ω the angular frequency and k the wave vector

$$E = \hbar\omega \tag{2}$$

$$P = \hbar k . \tag{3}$$

Figure 1 shows, as virtual holography-experiment, how the topological split-up due to the presence of a wire lead to interference of a single particle like an electron with itself.

A single electron, injected from the top, can pass the thin wire to the right and to the left. Accordingly the entangled matter-wave Ψ underneath the wire can be written as

$$X\Psi = X\left(\Psi_r + \Psi_l\right) \tag{4}$$

where X is a symbol to denote the strong entanglement due to mass conservation.

The probability to find a punctual interaction of the mass entangled matter-wave with the screen at point x is equivalent to the square of the matter-wave. Due to the fact that Ψ_r, Ψ_l differ in their path and phase, the phenomenon that a single electron can interfere with itself, occur. In case of an interaction the entire entangled matter wave interact as a whole to preserve mass conservation. This is the well known wave-particle duality.

Entangled matter-waves are the most appropriate representation for masses, where the particulate characteristic is caused by mass conservation. Correspondingly a single photon is composed of energy entangled waves with a continuous energy and direction distribution, thus a single photon can also interfere with itself, showing wave-particle duality.

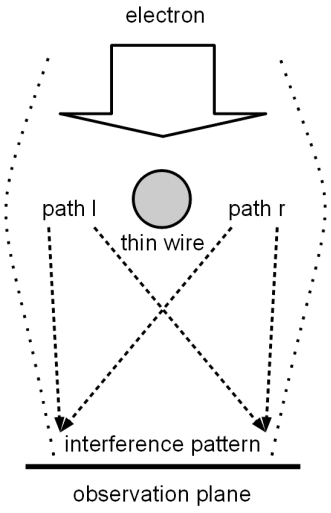


Figure 1. Interference between matter-waves of a single particle like an electron.

2.2. Spin Entanglement

In spacious systems the spin entanglement induces a spooky distant effect [3] [4]. If, for example, a two atomic molecule gets dissociated by an excitation (**Figure 2**), the atoms may remain entangled by their spins. If we measure the spin of both Atoms, one will show spin +1/2 and the other −1/2, but it is not possible to predict which one will have the positive and negative sign.

If we carefully align the spin of one atom in a magnetic field, the other must be oriented simultaneously in the opposite direction to conserve the total spin of both atoms, even though it is not in the vicinity. In other words, you can play with Schrödinger's cat while it's in the box. This "alignment" doesn't act as a force, where the first spin turns the second around, only the information about which state has to interact is "transferred" to the second spin to preserve the total spin of the entangled system.

2.3. Thermal Entanglement

At 0 ˚K almost all matter is in the lowest energy-state possible which I call the coherent state. In this state a system must interact as a whole, for example a solid shows a perfect Mößbauer effect or a fluid becomes a superfluid. Increasing the temperature result in distortions of the system and part of the system lose their coherence due to thermal chaos, where only the overall entanglement stays the same. The system gets split into coherent and normal phases where the length of coherence indicates the average size of coherent areas.

3. Applications

3.1. Electrical Superconductivity

The electrons in an atom are fragmented into paired wave-functions (orbitals) of constant energy but antipodal spin (entangled spins), which induces a partly bosonic characteristic [5]. This is why I would like to call these orbitals bound Cooper-pairs. In a solid these orbitals are shared with neighboring atoms (directed binding e.g. ceramics) or more spread out (metallic binding).

At 0 ˚K the orbitals of all atoms in a flawless crystal are perfectly spin entangled and the width of an energy band is reduced to a single value (BEC ground state for this energy band). The material is in the superconducting state.

Increasing the temperature result in distortions of the lattice and part of the orbitals lose their entanglement (bosonic characteristic) due to thermal chaos, where only the overall spin stays the same. The body gets split into coherent and solid phases. Below a critical temperature T_c the coherent phases are interconnected, forming a percolating superconducting backbone. Above T_c the density of the coherent phases is to low to form up a percolating backbone and the body is in the normal conducting state.

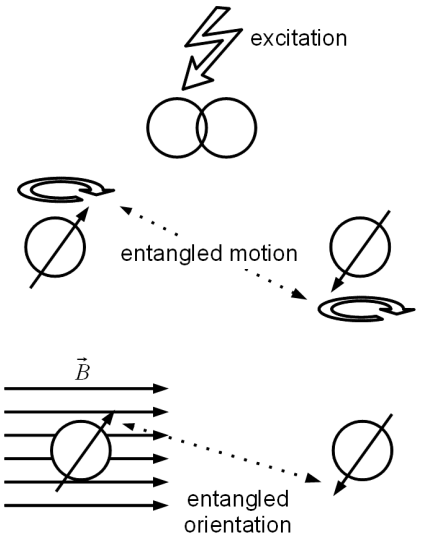


Figure 2. Spin entanglement between two atoms.

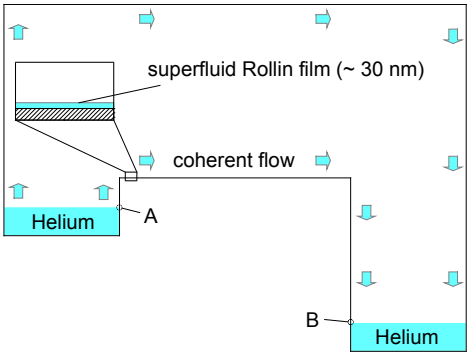


Figure 3. Entangled flow of superfluid Helium.

3.2. Superfluidity

The effect of superfluidity was first discovered in 1937 by Pyotr Kapitsa, John F. Allen and Don Misener at temperatures as low as 2.2 ˚K [6] [7]. In this variety of matter the thermally entangled coherent state of aggregation has zero viscosity and can show different quantum mechanical behavior on large scales. For example liquid helium can creep up walls and barriers to take up an energetically favorable state (**Figure 3**).

In a cavity superfluid Helium can flow over a barrier into a second basin with a lower gravitational potential. Responsible for this phenomenon is a thin superfluid layer (Rollin film) covering everything within the cavity. In the entangled state of aggregation an atom at A is thermally entangled with an atom at B. Both can only move the same way (coherent motion) and analog to the principle of corresponding pipes a flow takes place toward the basin with the lower gravitational potential.

3.3. Ultracold Gases

Ultracold gases are produced by a sequence of different cooling steps. Usually a couple of atoms are cooled using a laser. These atoms can be caught in a magneto optical trap. To reach the lowest possible temperature of the gas, the trap is adjusted so that the atoms with the highest temperature can evaporate out of the gas [8].

Depending on the temperature where the aggregation takes place, different gases can be generated.

When the temperature is high enough, the entanglements of the particles can be transferred to the gas. This gas is able to form a BEC.

If the aggregation temperature of the gas is to low to break the long ranged entanglements of the particles with its source, no BEC can be formed. The gas behaves like a perfect fermionic gas.

4. Conclusions

Some scientists may say that a quantum mechanical system is defined by the states their elements can reach and after assigning occupation probabilities they know everything about it. I hope I am able to show you that this is not in general the case. I don't believe that a quantum mechanical system where entanglement is present can be separated into elemental parts, where their fundamental properties like the entropy can be simply added up. Only the system as a whole defines measurable states. In general the system is in all these states simultaneously and the entanglement forces that a measurement gives a value for one state. A more fanciful opinion may say that Schrödinger's cat exists in an entangled quantum-multiverse until it is forced to interact at which the entanglement defines a single universe. With respect to the observed pattern in **Figure 1**, this illustrative model may be completed by adding the capability of interference in the space of the quantum-multiverse. Accordingly the symbol X in Equation (4) denotes that the state Ψ is not an element of an Hilbert space but a state representing a quantum-multiverse.

I wonder if it is possible to understand for example the exchange of virtual photons in quantum-electrodynamics (QED) as information exchange in an entanglement driven sense. Due to an interaction (in QED described by the exchange of a virtual photon), the quantum-multiverse is forced to define a single universe. With this in mind, Quantum Field Theories provide a suitable mathematical abstraction for the physical reality of quantum entanglement.

References

[1] Einstein, A. (1905) Zur Elektrodynamik bewegter Körper. *Annalen der Physik und Chemie*, **17**, 891-921. http://dx.doi.org/10.1002/andp.19053221004

[2] de Broglie, L. (1929) The Wave Nature of the Electron. *Nobel Lecture*, 12.

[3] Einstein, A., Podolsky, B. and Rosen, N. (1935) Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, **47**, 777-780. http://dx.doi.org/10.1103/PhysRev.47.777

[4] Afriat, A. and Selleri, F. (1999) The Einstein, Podolsky, and Rosen Paradox in Atomic, Nuclear, and Particle Physics. Plenum Press, New York. http://dx.doi.org/10.1007/978-1-4899-0254-2

[5] Cordelair, J. (2014) Superconductivity. *World Journal of Condensed Matter Physics*, **4**, 241-242. http://dx.doi.org/10.4236/wjcmp.2014.44026

[6] Kapitza, P. (1938) Viscosity of Liquid Helium below the λ-Point. *Nature*, **141**, 74. http://dx.doi.org/10.1038/141074a0

[7] Fairbank, H.A. and Lane, C.T. (1949) Rollin Film Rates in Liquid Helium. *Physical Review*, **76**, 1209-1211. http://dx.doi.org/10.1103/PhysRev.76.1209

[8] Hau, L.V. (2001) Frozen Light. *Scientific American*, **285**, 52-59. http://dx.doi.org/10.1038/scientificamerican0701-66

PAPER • **OPEN ACCESS**

Detecting quantum entanglement with unsupervised learning

To cite this article: Yiwei Chen *et al* 2022 *Quantum Sci. Technol.* **7** 015005

Quantum Science and Technology

PAPER

Detecting quantum entanglement with unsupervised learning

Yiwei Chen[1], Yu Pan[1,*], Guofeng Zhang[2,3,*] and Shuming Cheng[4,5,6,*]

[1] Institute of Cyber-Systems and Control, College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, People's Republic of China
[2] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong Special Administrative Region of China, People's Republic of China
[3] Shenzhen Research Institute, The Hong Kong Polytechnic University, Shenzhen 518057, People's Republic of China
[4] The Department of Control Science and Engineering, Tongji University, Shanghai 201804, People's Republic of China
[5] Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 201804, People's Republic of China
[6] Institute for Advanced Study, Tongji University, Shanghai, 200092, People's Republic of China
[*] Authors to whom any correspondence should be addressed.

E-mail: ypan@zju.edu.cn, Guofeng.Zhang@polyu.edu.hk and shuming_cheng@tongji.edu.cn

Keywords: entanglement detection, machine learning, quantum resource

OPEN ACCESS

RECEIVED
19 June 2021

REVISED
8 October 2021

ACCEPTED FOR PUBLICATION
19 October 2021

PUBLISHED
3 November 2021

Abstract

Quantum properties, such as entanglement and coherence, are indispensable resources in various quantum information processing tasks. However, there still lacks an efficient and scalable way to detecting these useful features especially for high-dimensional and multipartite quantum systems. In this work, we exploit the convexity of samples without the desired quantum features and design an unsupervised machine learning method to detect the presence of such features as anomalies. Particularly, in the context of entanglement detection, we propose a complex-valued neural network composed of pseudo-siamese network and generative adversarial net, and then train it with only separable states to construct non-linear witnesses for entanglement. It is shown via numerical examples, ranging from two-qubit to ten-qubit systems, that our network is able to achieve high detection accuracy which is above 97.5% on average. Moreover, it is capable of revealing rich structures of entanglement, such as partial entanglement among subsystems. Our results are readily applicable to the detection of other quantum resources such as Bell nonlocality and steerability, and thus our work could provide a powerful tool to extract quantum features hidden in multipartite quantum data.

1. Introduction

Peculiar quantum features, signalled by quantum entanglement [1] and coherence [2], enable us to accomplish tasks impossible for classical systems [3], such as ensuring the security of communications and speeding up certain hard computational tasks [4, 5]. Hence, an important question naturally arises: how can the presence of these features be efficiently detected for any given quantum system? Indeed, this is a challenging task for high-dimensional and multipartite systems because quantum features usually imply correlated patterns hidden within subsystems. Taking entanglement for example, except for low-dimensional systems, e.g. $2 \otimes 2$ and $2 \otimes 3$, of which entanglement could be detected faithfully via the positive partial transpose (PPT) criterion [6], generically, it is an NP-hard problem [7]. Besides, even though at least one linear entanglement witness could be found to witness any entangled state [1, 8–10] as displayed in figure 1, there still lacks a universal and scalable way to construct such an appropriate witness for an arbitrary state in practice.

In this work, we turn to the machine learning technique which is powerful in extracting features or patterns hidden in large multipartite datasets to tackle the quantum detection problem. Recently, much progress has been achieved in this inter-disciplinary field of quantum machine learning [11]. For example, on one hand, many quantum or quantum-inspired algorithms have been developed to speed up some well-known machine learning algorithms [12–14]. On the other hand, machine learning is also a natural candidate to extract correlated features of high-dimensional quantum systems, which has found wide

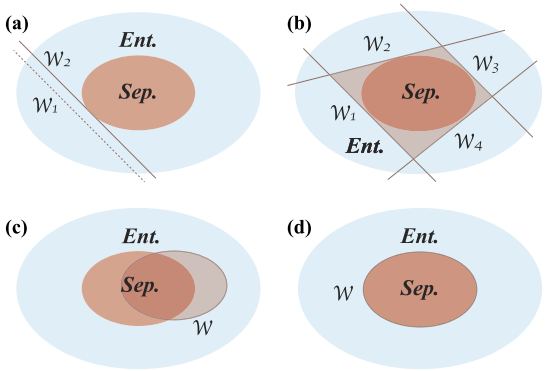


Figure 1. The geometrical interpretation of entanglement detection via entanglement witnesses. (a) Standard linear entanglement witnesses. The witness $\mathcal{W}_2$ is finer than $\mathcal{W}_1$. (b) A proper set of linear entanglement witnesses is able to form a closed area which encloses all separable states. Here $\mathcal{W}_1, \ldots, \mathcal{W}_4$ are used as examples. (c) An imperfect nonlinear entanglement witness could be generated via supervised learning method if we cannot label enough samples to cover the space of entangled states. (d) A near-perfect nonlinear witness $\mathcal{W}$ can be approximately constructed by the unsupervised neural network if the generated training data span the space of separable states.

applications in quantum control [15], state tomography [16], measurement [17, 18], and many-body problems [19–21]. Especially, the task of quantum entanglement detection can be formulated as a binary classification problem. As a consequence, various classical neural nets, trained with both entangled and separable samples, have been constructed to solve this problem via supervised learning [22–24]. However, the supervised training method requires a large pre-labelled dataset. In practice, it is time-consuming or even impossible to faithfully label a large number of entangled states in a high-dimensional space [7], thus leading these supervised methods into a dilemma.

Here, we instead build up an unsupervised model to accomplish the task of entanglement detection beyond the above issues. Following from the fact that separable states form a convex set, it becomes an anomaly detection problem of which all separable samples are labelled as normal and entangled ones are abnormal. Particularly, as shown in figure 3, a class of complex-valued neural networks composed of a pseudo-siamese network and a generative adversarial net (GAN), is constructed and then trained with very few normal samples to detect entanglement for multipartite systems, ranging from two-qubit to ten-qubit states. It is noted that our model is much more feasible than anomaly detection methods proposed in [25, 26] which require quantum hardware.

It is further illustrated in figure 1 that our unsupervised neural nets are essentially trained to search for proper nonlinear entanglement witnesses which near-perfectly construct the boundary between separable and entangled samples. Numerical results show that it is able to achieve extremely high accuracy of entanglement detection with above 97.5% on average, and even capable to detect partial entanglement within subsystems, e.g. bi-separable states in three-qubit system with accuracy above 97.7%.

Our work is organised as follows. In section 2, we give a brief introduction to the task of entanglement detection and unsupervised learning method. Then we propose an unsupervised learning neural network targeted for the detection of generic quantum features. In section 3, multipartite entanglement detection is taken as examples to illustrate the performance of our model, with only separable samples used for training. Finally, we conclude this work with a summary in section 4.

2. Unsupervised entanglement detection

2.1. The task of detecting entanglement

Entanglement is not only of significant importance to understand quantum theory at the fundamental level [1], but also has found applications in information protocols, such as quantum teleportation [27]. For a given n-partite quantum system, entanglement associated with the state is defined in a passive way in which a state ρ is entangled if and only if it cannot be described in a fully-separable form of [28]

$$\rho_{\text{sep}} = \sum_{i=1}^{m} \lambda_i \rho_i^1 \otimes \cdots \otimes \rho_i^j \otimes \cdots \otimes \rho_i^n \tag{1}$$

with non-negative coefficients satisfying $\sum_{i=1}^{m} \lambda_i = 1$. Here ρ_i^j denotes the state density matrix of the jth subsystem. Obviously, all of the separable states as per equation (1) form a convex set in the sense that any convex combination of these states in this set also belong to the same state set. It is noted that the above

definition of entanglement does not fully capture the entangled structure in the state, e.g. the partial entanglement [29], which will be discussed later.

In practice, whether a given state ρ is entangled or not, can be experimental-friendly determined via an entanglement witness [1, 10]. Indeed, as shown in figure 1(a), an entanglement witness essentially defines a hyperplane which separates the entangled state from the convex set of separable states. Furthermore, it has been shown in [1] that it is impossible for one linear witness to detect all entangled states, implying that a large set of linear witnesses illustrated in figure 1(b) (could be impractical) or certain nonlinear witness shown in figure 1(d) may be required. Besides, it becomes extremely inefficient and impractical to construct a proper witness for an arbitrary state, especially in multipartite systems. The entanglement witnesses as neural networks are experimentally accessible and has been demonstrated in [24]. In fact, since neural networks are learning the linear and nonlinear correlations on the quantum states to form a classifier, a properly parameterized neural network layer is equivalent to a set of generalized Bell's inequalities for the experimental detection of entanglement. In the following, we propose a complex-valued neural network trained in unsupervised manner to search for the nonlinear entanglement witnesses as desired.

2.2. Unsupervised learning

The unsupervised model refers to the process of learning a probability distribution over the data that has not been classified or categorized. In this situation, automated methods or algorithms must explore the underlying features from the available data and group them with similar characteristics. Specifically, the unsupervised model only receives a training set $\mathcal{S}$ that contains

$$\mathcal{S} = \{x_1, x_2, x_3, \ldots\} \tag{2}$$

without supervised target outputs $\{y_1, y_2, y_3, \ldots\}$. In contrast to supervised learning where tagging data requires a large amount of time, unsupervised learning exhibits high efficiency and self-organization in capturing patterns from untagged data.

Autoencoder [30] is a widely used unsupervised learning method that aims to learn efficient representations for a set of data. Typically, an autoencoder consists of two modules, namely encoder $\mathcal{E}$ and decoder $\mathcal{D}$, where the former learns the latent representation (encoding) for input data, and the latter is trained to generate an output as close as possible to its original input from the latent representation. Another well-known unsupervised learning method is GAN [31, 32]. Specifically, two neural networks, namely generator $\mathcal{G}$ and discriminator $\mathcal{D}$, contest with each other in the form of a zero-sum game in GAN, where the gain of one module is the loss of the other. This technique learns to generate new data with the same statistics as the training set. The siamese network [33, 34], as shown in figure 2, contains a pair of neural networks built by the same parameters, which receives two inputs and detects their difference by comparing the output vectors of the networks. The siamese network is capable of learning generic features for making predictions about an unknown distribution even when few examples from the distribution are available, which provides a competitive approach for pattern recognition without the domain-specific knowledge. In particular, the siamese network can be trained in an unsupervised manner, as the labels of the input data are not needed.

For these reasons, the method proposed in this paper has been built upon the siamese network, which is suitable for one-class unsupervised learning. The basic idea is similar to one-class support vector machine for anomaly detection [35]. That is, given a set of training samples, we aim to model the underlying distribution of the data and detect the soft boundary of this set, in order to classify new inputs as belonging to this set or not. In this case, the model will only take a training dataset without class labels as input, which means the model is a type of unsupervised learning methods.

2.3. Constructing the complex-valued neural networks

As shown in figure 3, our networks could be decomposed into two parts: one is the pseudo-siamese neural network (in the red dashed box) and the other is the GAN (in the purple dashed box). The complex-valued neural network receives the density state matrix as the input. The building modules for these networks are detailed in appendix A.

The pseudo-siamese neural network consists of two encoders sharing the same network structure, labelled as $\mathcal{E}_\mathrm{r}$ and $\mathcal{E}_\mathrm{g}$, respectively. In contrast to the original siamese network [34] which requires quadratic pairs as input, the pseudo-siamese network only requires a single input ρ_real be fed to the first encoder $\mathcal{E}_\mathrm{r}$. The second input ρ_gen to the second encoder $\mathcal{E}_\mathrm{g}$ is automatically generated by the decoder $\mathcal{G}$ whose aim is to reconstruct ρ_real. Therefore, the pseudo-siamese network trains much faster than the original siamese network while inherits its few-shot learning ability. In principle, these two encoders competes with each other to produce a pair of indistinguishable feature vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. The performance is evaluated by the

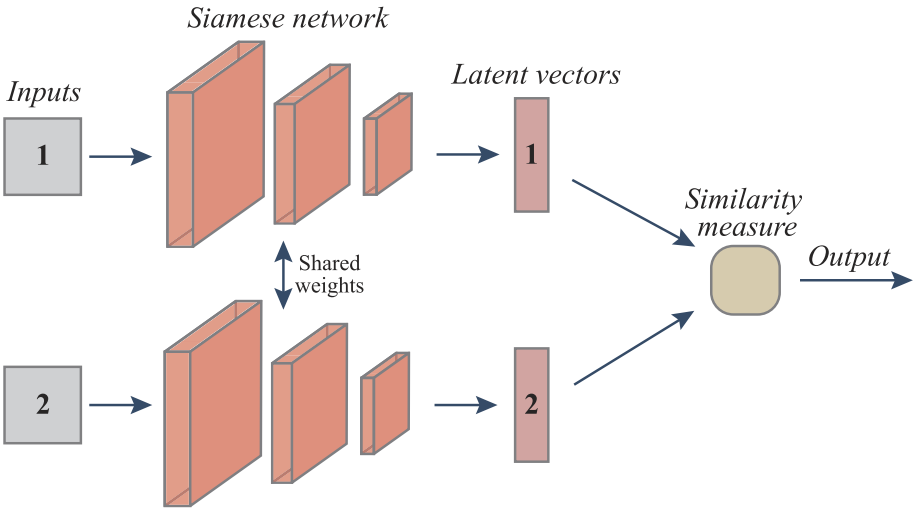


Figure 2. Pipeline of the siamese network. Two inputs are encoded into latent vectors whose difference is detected by a similarity measure.

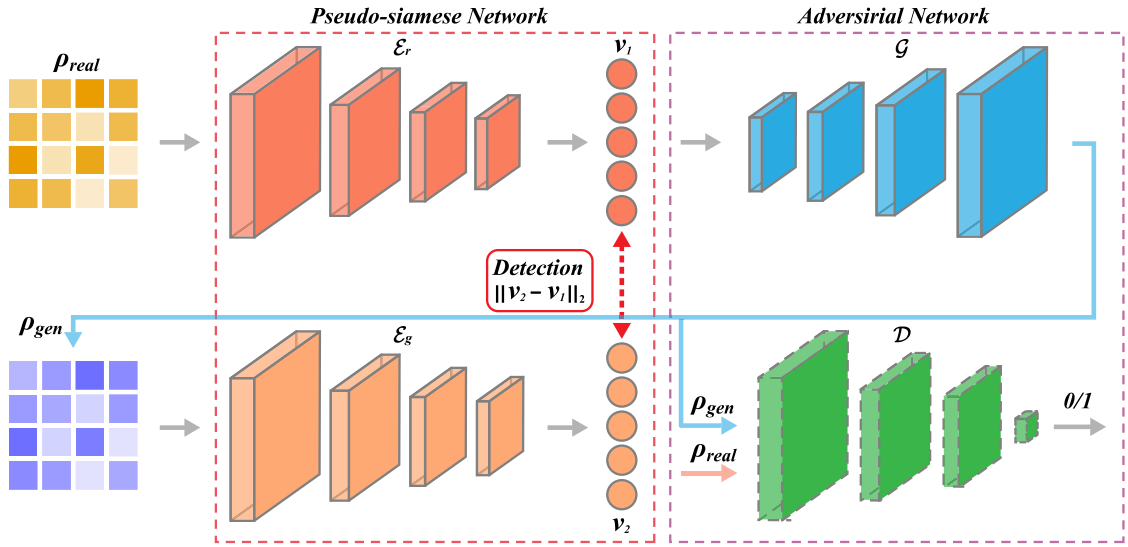


Figure 3. Structure of the complex-valued neural network. The complex-valued network is composed of two parts: one is the pseudo-siamese network in the red dashed box and the other is a GAN in the purple dashed box. The pseudo-siamese neural network consists of two encoders $\mathcal{E}_r$ and $\mathcal{E}_g$ that share the same network structure. ρ_{gen} is generated by $\mathcal{G}$. The discriminator network $\mathcal{D}$ is a binary classifier which outputs either 0 or 1. The generator $\mathcal{G}$ and discriminator $\mathcal{D}$ form the GAN, which aims to produce a ρ_{gen} that is as close as possible to ρ_{real}.

cost function

$$\mathcal{L}_1 = \mathbf{E}_{\rho_{real}} \|\mathcal{E}_r(\rho_{real}) - \mathcal{E}_g(\mathcal{G}(\mathcal{E}_r(\rho_{real})))\| = \mathbf{E}_{\rho_{real}} \|v_1 - v_2\|, \tag{3}$$

where the norm $\|x\|$ could be the L_p-norm of any complex vector x with $\|x\|_p \equiv (|\Re(x)|^p + |\Im(x)|^p)^{1/p}$. Here two-norm is chosen for equation (3). As the two inputs to the encoders $\mathcal{E}_r$ and $\mathcal{E}_g$ are slightly different, the two encoders would not share the same weight parameters after training [36].

Combining the encoder $\mathcal{E}_r$ with $\mathcal{G}$ yields an encoder–decoder structure which aims to produce fake samples that are close to real ones. Thus we introduce the loss function

$$\mathcal{L}_2 = \mathbf{E}_{\rho_{real}} \|\rho_{real} - \mathcal{G}(\mathcal{E}_r(\rho_{real}))\| = \mathbf{E}_{\rho_{real}} \|\rho_{real} - \rho_{gen}\| \tag{4}$$

to quantify its performance. In analogy to classical autoencoders [37], it is found that L_1-norm achieves better performance than that of $p = 2$ for this loss term.

An optional discriminator network could be introduced for additional adversarial training. The discriminator $\mathcal{D}$ and generator $\mathcal{G}$ form the GAN (figure 3) which could enhance the ability of $\mathcal{G}$ to produce more realistic quantum samples. Indeed, $\mathcal{D}$ is a binary classifier trained to discriminate fake samples from

real ones. The two cost functions for this adversarial net are given by

$$\mathcal{L}_{\text{adv1}} = \mathbf{E}_{\rho_{\text{real}}}(-\mathcal{D}(\rho_{\text{real}}) + \mathcal{D}(\mathcal{G}(\mathcal{E}_{\text{r}}(\rho_{\text{real}})))), \tag{5}$$

$$\mathcal{L}_{\text{adv2}} = \mathbf{E}_{\rho_{\text{real}}}(-\mathcal{D}(\mathcal{G}(\mathcal{E}_{\text{r}}(\rho_{\text{real}})))), \tag{6}$$

which are alternatively minimized via gradient descent method. Specifically, the gradients are clipped between -1 and 1, turning the network into a Wasserstein GAN which is easy to train [32]. In each round, the parameters of $\mathcal{D}$ are updated by minimizing $\mathcal{L}_{\text{adv1}}$, while the parameters of $\mathcal{G}$ and $\mathcal{E}_{\text{r}}$ are updated by minimizing $\mathcal{L}_{\text{adv2}}$.

Finally, by combining (3)–(6), the complex-valued neural network is trained by alternatively minimizing $\mathcal{L}_{\text{adv1}}$ and

$$\mathcal{L}_3 = w_1 \cdot \mathcal{L}_1 + w_2 \cdot \mathcal{L}_2 + w_a \cdot \mathcal{L}_{\text{adv2}}, \tag{7}$$

with the weight parameters w_1, w_2, and w_a being chosen adaptively.

2.4. Training the networks via unsupervised learning

Suppose the complex-valued network is trained with separable states only, an entangled state would result in a feature vector $\boldsymbol{v}_{\text{ent}}$ distinct from that of the generated one in the latent space. Indeed, the entire training and prediction process can be divided into three steps as follows.

(a) Preparing separable states as training samples. Following equation (1), each ρ_i^j is generated via $HH^\dagger/(\text{tr}\, HH^\dagger)$, where H is a complex-valued matrix whose real and imaginary parts of each entry are sampled from independent Gaussian distributions. It is noted that this sampling method could cover the whole space of separable states [38].

(b) Training the neural network on the generated set of separable states by alternatively minimizing $\mathcal{L}_{\text{adv1}}$ as per equation (5) and $\mathcal{L}_3$ as per equation (7) via the gradient descent method.

(c) Determining the decision threshold value b on the test set after training. We choose b to satisfy

$$\frac{\text{FN}}{\text{TP}+\text{FN}} = \frac{\text{FP}}{\text{FP}+\text{TN}}, \tag{8}$$

where TP, FP, TN, and FN refer to the number counts of true positive, false positive, true negative, and false negative samples. Here, being positive or negative stands for a separable or entangled sample. Choosing b to satisfy equation (8) implies that the probabilities of misclassifying entangled and separable states are the same on the test set. Hence, if the score of a quantum state is larger than this b, then it will be detected as entangled.

For each ρ in the test set, its score for entanglement detection can be defined as

$$\mathcal{A}(\rho) = \left\|\mathcal{E}_{\text{r}}(\rho) - \mathcal{E}_{\text{g}}(\mathcal{G}(\mathcal{E}_{\text{r}}(\rho)))\right\|_2. \tag{9}$$

It could be further expressed in a witness-like form of

$$\mathcal{A}(\rho) = \|(\mathcal{W}_{\mathcal{E}_{\text{g}}}\mathcal{W}_{\mathcal{G}} - \mathcal{I})\mathcal{W}_{\mathcal{E}_{\text{r}}} \cdot \text{vec}(\rho)\|_2 = \|\mathcal{W} \cdot \text{vec}(\rho)\|_2, \tag{10}$$

where $\mathcal{W}_{\mathcal{E}_{\text{f}}(\mathcal{G})}$ denotes the weight tensor which generates the corresponding linear and nonlinear network transformations. For this reason, the neural network model can be regarded as trying to determine the nonlinear witness $\mathcal{W}$ which approximately characterizes the boundary between separable and entangled states, without relying on samples of entangled states during training.

Alternately, there is another way to implement the model for prediction without the test dataset, making both training and prediction independent of any information of entangled states. This is achieved by determining b as

$$b = \max_{\rho_{\text{sep}}} \mathcal{A}(\rho). \tag{11}$$

Obviously, this approach leads to a higher detection accuracy than using equation (8). Since both the training and implementation do not rely on entangled samples, this approach is computationally efficient. More importantly, the major advantage of our unsupervised learning framework lies in its scalability, as generating sufficient entangled states for training becomes impractical for high-dimensional quantum systems.

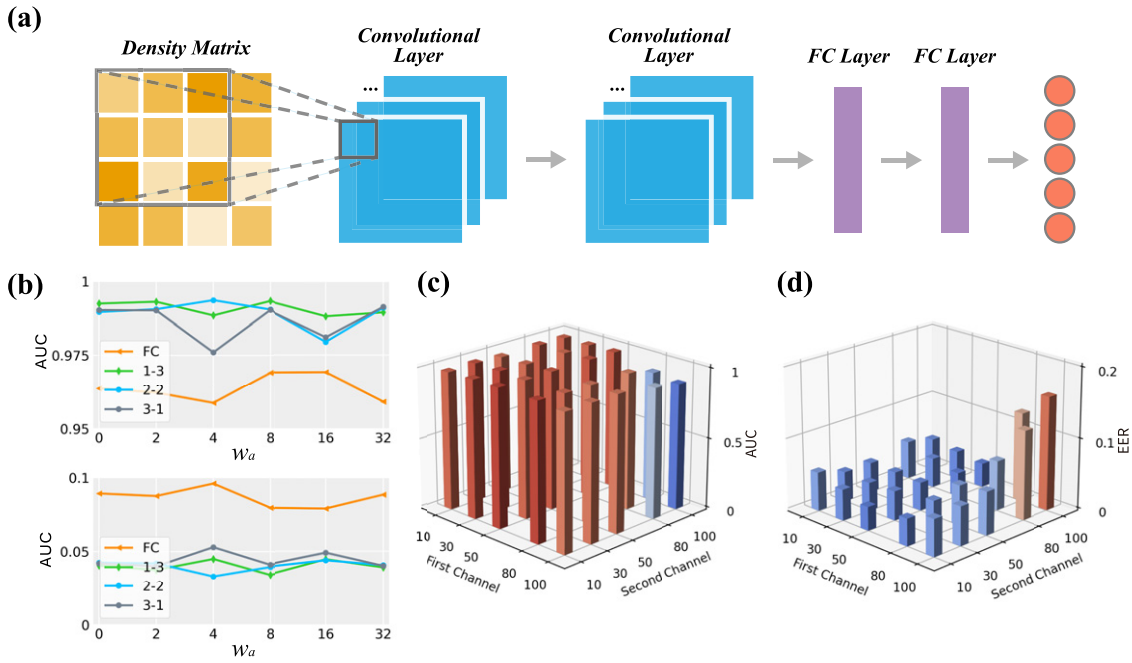


Figure 4. Detection of two-qubit states. (a) Four-layer neural network structure of the encoder, including two convolutional layers followed by two FC layers. The structures of the encoder and decoder are symmetric. The structure of the discriminator is the same as the encoder, with an additional normalization layer to produce a scalar output. (b) The performance of the neural networks with the first two layers being convolutional or FC. The convolutional kernel size combinations that have been tested are 1×1 and 3×3, 2×2 and 2×2, 3×3 and 1×1, with the number of output channels being 10 and 30, respectively. If the first two layers are FC, the number of output channels is set to 32 and 128, respectively. (e) and (f) AUCs and EERs with different number of output channels for convolutional layers. The convolutional kernel sizes are 2×2 and 2×2.

3. Numerical results

3.1. Evaluation metrics

We use two evaluation metrics of binary classification in our experiments. The first metric is the area under curve (AUC) of the receiver operating characteristic curve, which is created by plotting the true positive rate (TPR $=$ TP/(TP $+$ FN)) against the false positive rate (FPR $=$ FP/(FP $+$ TN)) using the similarity score defined in (9) for various values of b [39]. The second metric is equal error rate (EER), which is defined as FN/(TP $+$ FN) when equation (8) holds [40].

3.2. Detecting two-qubit entangled states

The number of training samples for two-qubit case is 160 000, all composed of separable states. The number of testing samples is 80 000, including 40 000 separable states and 40 000 entangled states. Two-qubit separable states are generated by

$$\rho_{\text{sep}} = \sum_{i=1}^{m} \lambda_i \rho_i^1 \otimes \rho_i^2, \tag{12}$$

where $\sum_{i=1}^{m} \lambda_i = 1$ and $0 \leqslant \lambda_i \leqslant 1$, with m iterating from 1 to 20. Entangled states are selected from randomly generated states of the entire system using PPT criterion.

The structure of the four-layer encoder is illustrated in figure 4(a). The last two layers of the encoder are fully-connected (FC) layers, with output channels being 64 and 10, respectively. The first two layers can be convolutional with different kernels and different number of output channels, or fully connected as tested in figure 4(b). The best performance of the model has been achieved with the convolutional kernel size of the first two layers being 2×2 and 2×2. The best AUC is 0.99 and EER is 2.99%, attained at a small w_a which is the weight of adversarial cost for training. As shown in figure 4(c), convolutional layer performs much better than FC layer, with AUC being consistently higher than 0.975 and EER lower than 5%. Figures 4(e) and (f) shows the performance of convolutional neural networks when the number of output channels varies, indicating that a small number of output channels is enough to extract the features of entanglement for two-qubit states.

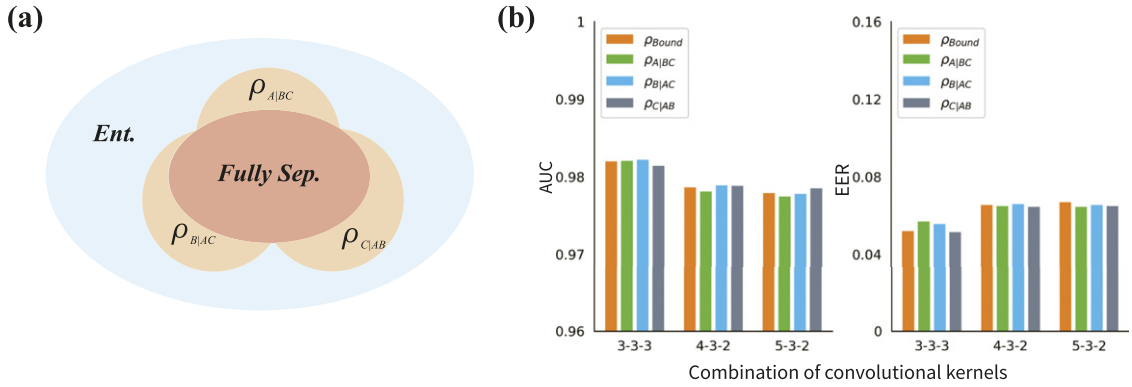


Figure 5. Detection of three-qubit states. (a) Distribution of three-qubit states. $\rho_{A|BC}$ is a bi-separable state with qubit B and C entangled. (b) EERs and AUCs with different combinations of convolutional kernels, where i–j–k stands for the kernel sizes of the three convolutional layers.

3.3. Detecting three-qubit entangled states

An entangled three-qubit state can be classified into several types, e.g. bi-separable states and bound entangled states [24]. The three-qubit state is fully-separable if

$$\rho_{\text{sep}} = \sum_{i=1}^{m} \lambda_i \rho_i^{\text{A}} \otimes \rho_i^{\text{B}} \otimes \rho_i^{\text{C}}. \tag{13}$$

The distribution of three-qubit states is illustrated in figure 5(a). In this case, successful supervised learning requires that one can generate enough and balanced samples for all types of entanglement, which cannot be guaranteed by the current random sampling techniques. In contrast, a universal entanglement detector could be built using only the fully-separable samples if unsupervised learning method is employed.

The numerical results in figure 5(b) are based on a dataset consisting of 160 000 training samples and 200 000 test samples. The training samples are fully-separable states, and the test samples include 40 000 fully-separable states, 40 000 bound entangled states and 120 000 bi-separable states (40 000 for each subtype). To accommodate the 8×8 density matrix input, a third convolutional layer is added. The number of the output channels for the three convolution layers is 10, 30, 50, respectively. Since the unsupervised model focuses on detecting the feature of separability instead of the features of different types of entanglement, it has achieved similar detection accuracy on four types of entangled samples.

The proposed unsupervised learning method is applicable to the detection of partial entanglement and genuine entanglement. Here we take the detection of bi-separable states of a three-qubit system as an example [41]. Suppose the task is to discriminate the bi-separable states $\rho_{\text{A|BC}}$ (B and C are entangled) from the other states. By generating the entangled states for subsystem BC using the PPT criterion, the samples of bi-separable states are given by

$$\rho_{\text{A|BC}} = \sum_{i=1}^{m} \lambda_i \rho_i^{\text{A}} \otimes \rho_i^{\text{BC}}. \tag{14}$$

A classifier for A|BC separability can be obtained by training on these samples in an unsupervised manner. Particularly, if we replace ρ_i^{BC} in (14) by a generic two-qubit state, the anomalies detected would be the quantum states that are entangled between A and BC (page 10). Furthermore, if we generate the samples as

$$\rho_{\text{ABC}} = \sum_{i=1} \lambda_i^1 \rho_i^{\text{A}} \otimes \rho_i^{\text{BC}} + \sum_{j=1} \lambda_j^2 \rho_j^{\text{B}} \otimes \rho_j^{\text{AC}} + \sum_{k=1} \lambda_k^3 \rho_k^{\text{C}} \otimes \rho_k^{\text{AB}}, \tag{15}$$

the abnormal samples detected by the unsupervised model would be quantum states which are not bi-separable. In other words, the genuine entanglement of the three-qubit state can be detected as an anomaly.

3.4. Scalability up to ten-qubit states

The unsupervised learning method is applied on four- to ten-qubit states to study its scalability. We have found that the generation of separable states for training is very efficient even for tens of qubits, because the generation of separable pure states is very efficient, which is done by generating single qubit states and calculating their Kronecker products. Consequently, mixed (fully and partial) separable states can be constructed as linear combinations of pure states, which does not take much time. In this work, it takes less

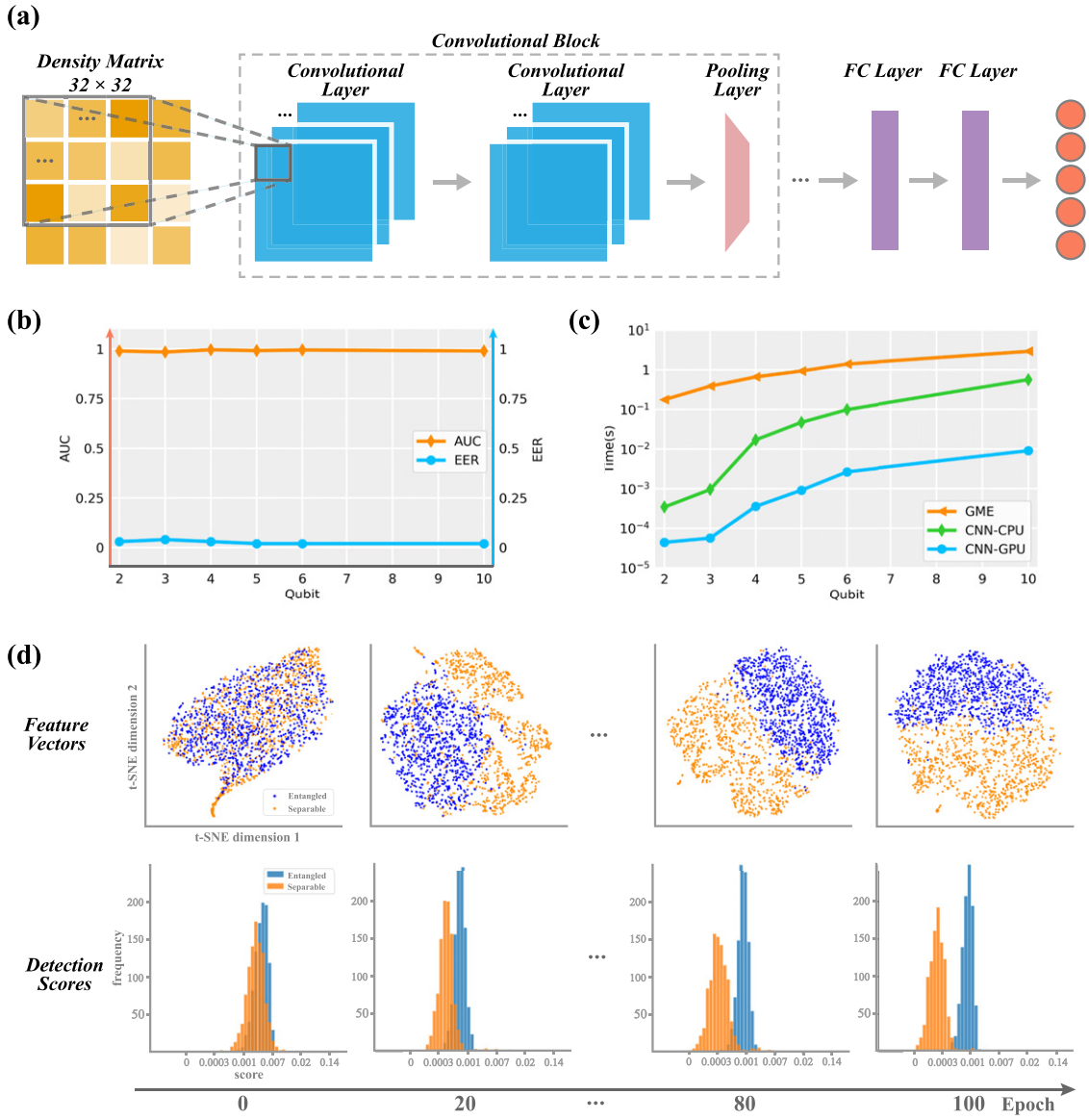


Figure 6. Scalability of unsupervised learning. (a) The neural network structure of the encoder for five-qubit pure states. The convolutional block is composed of two convolutional layers and a 2 × 2 max pooling layer. (b) AUCs and EERs achieved by the unsupervised learning method as the number of qubits increases. (c) The comparison of inference time between the up-to-date numerical method for computing the GME and the neural network method. (d) The evolution of feature vectors and detection scores for 1000 separable and 1000 entangled samples of five-qubit states during training.

than 10 min to generate enough pure separable samples for ten-qubit states on a desktop computer, and mixing the samples takes less than 3 min. Moreover, we have observed a linear increase on the generation time with the dimension. Here we used pure four- to ten-qubit states for training because the test samples of mixed states (mixed entangled states) are hard to label for high-dimensional system, while we have developed an efficient algorithm [42] that can tell whether a randomly generated ten-qubit pure state is entangled or not within 5 s. However, test samples are just used to measure the accuracy of the model. The model is trained using the separable samples only, which can be generated efficiently. The trained model can be implemented without using test samples as shown in (11). Therefore, the model can also be trained and implemented with mixed state samples for high-dimensional cases. Note that the geometrical measure is only used to label the entangled states for the test dataset. The separable pure states are generated by

$$|\psi_{\text{sep}}\rangle = |\psi_i^1\rangle \otimes \cdots |\psi_i^j\rangle \cdots \otimes |\psi_i^n\rangle, \tag{16}$$

where $|\psi_i^j\rangle$ is a randomly generated pure state vector of the jth qubit. The real and imaginary parts of the complex-valued vector are sampled from an independent Gaussian distribution. The density matrix $\rho_{\text{sep}} = |\psi_{\text{sep}}\rangle\langle\psi_{\text{sep}}|$ is used as the input to the neural network. Figure 6(a) depicts the network structure of the encoder for entanglement detection in five-qubit states, where a max pooling layer has been added to

handle the increased dimension of the input. For ten-qubit states, we adopt three convolutional layers and increase the max pooling size to 4×4. The training dataset is composed of 160 000 separable states, and the test dataset is composed of 40 000 separable and 40 000 entangled states. The entangled states are found by randomly generating four- to ten-qubit pure states and computing their entanglement measures using the numerical method from [42]. See appendix B for the details of the algorithm.

As shown in figure 6(b), the unsupervised model achieves an AUC of 0.9952 and an EER of 2.02% for entanglement detection in ten-qubit states. The EER is 0.54% for entanglement detection in five-qubit states, which means only 54 in 10 000 states are misclassified. The short inference time is another advantage of the neural network model. The inference time of the neural network model on GPU is about tens of microseconds to hundreds of microseconds for up to 10 qubits (figure 6(c)), which is significantly faster than the up-to-date numerical method which takes the state vector instead of density matrix as the input for computing the geometrical entanglement measure (GME). The time needed for generating training dataset is greatly reduced as compared to supervised learning methods, since there is no need to label the entangled states. For example, suppose the ten-qubit training dataset of the supervised method consists of 100 000 samples, which must be labelled by numerically computing the GME. The total time needed for generating the dataset is about 138 h (labelling each sample takes 5 s in average). In contrast, generating separable training samples of the ten-qubit system is much more simple, which only takes several minutes.

The upper half of figure 6(d) shows the evolution of feature vectors of 1000 separable and 1000 entangled states in the training process for five-qubit states. We visualize the evolution by t-SNE method [43] which maps the feature vectors to two-dimensional space. In the first 10 epochs, the entangled and separable states are mixed up in the latent space and difficult to distinguish. After 20 epochs, the feature vectors start to split into two set. In the last 20 epochs, the feature vectors of separable states are separated completely from the feature vectors of entangled states, with very few exceptions. A similar evolution can be seen in the distribution of detection scores of the input states. After training, the detection scores of separable states are more closed to zero, while the scores of entangled states are concentrated around 0.001.

4. Conclusions and discussions

We have proposed an efficient and scalable method with unsupervised learning to detect quantum entanglement. Specifically, we build up a class of complex-valued pseudo-siamese neural networks which is easy to implement as it is trained without entangled samples. Moreover, it is scalable to detect entanglement of multipartite systems where sufficient labelled entangled samples become difficult to obtain, and our numerical analysis finds that we could still obtain a rather high accuracy with above 97.5% on average for multipartite systems from two-qubit to ten-qubit. For this reason, we believe that our work provides a promising tool to detect quantum features of high-dimensional quantum data.

Finally, it is noted that we exploit the convexity of separable samples and thus reformulate entanglement detection as an anomaly detection problem, for which the unsupervised neural networks are suitable. Since other useful quantum features, such as Bell nonlocality and Einstein–Podolsky–Rosen steerability, also share the same property that it is defined as a distinguishable sample from a convex set, it is evident that our work can be readily generalized to solve the similar detection problem.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grants Nos. 62173296 and 62088101. G F also acknowledges support from Hong Kong Research Grant Council (Grants Nos. 1520841, 15203619, and 15506619), Shenzhen Fundamental Research Fund, China, under Grant No. JCYJ20190813165207290, and the CAS AMSS-polyU Joint Laboratory of Applied Mathematics.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Appendix A. Complex-valued neural network

We build the complex-valued neural network based on the work of [44]. The codes are available at https://github.com/ewellchen/Entanglement_detection. The two-dimensional convolutional (denoted as *)

and FC (denoted as $\cdot$) operations of the weight w and input z in the complex domain are defined by

$$w * z = \Re\{w\} * \Re\{z\} - \Im\{w\} * \Im\{z\} \tag{A.1}$$

$$+ \mathrm{i}(\Im\{w\} * \Re\{z\} + \Re\{w\} * \Im\{z\}), \tag{A.2}$$

$$w \cdot z = \Re\{w\}\Re\{z\} - \Im\{w\}\ \Im\{z\} \tag{A.3}$$

$$+ \mathrm{i}(\Re\{w\}\Im\{z\} + \Im\{w\}\ \Re\{z\}), \tag{A.4}$$

where $\Re$ and $\Im$ represent the real and imaginary part of the vector or matrix, respectively. The formulation of the complex-valued rectified linear unit (CReLU) is given by

$$\mathrm{CReLU}(z) = \mathrm{ReLU}(\Re(z)) + \mathrm{iReLU}(\Im(z)), \tag{A.5}$$

which introduces nonlinearity into the network transformation. The batch normalization (BN) layer is implemented by multiplying the 0-centered data $(\boldsymbol{z} - \mathbf{E}[\boldsymbol{z}])$ with the inverse square root of the covariance as

$$\mathcal{V} = \begin{pmatrix} \mathrm{Cov}(\Re\{\boldsymbol{z}\}, \Re\{\boldsymbol{z}\}) & \mathrm{Cov}(\Re\{\boldsymbol{z}\}, \Im\{\boldsymbol{z}\}) \\ \mathrm{Cov}(\Im\{\boldsymbol{z}\}, \Re\{\boldsymbol{z}\}) & \mathrm{Cov}(\Im\{\boldsymbol{z}\}, \Im\{\boldsymbol{z}\}) \end{pmatrix},$$

$$\tilde{\boldsymbol{z}} = (\mathcal{V})^{-\frac{1}{2}}(\boldsymbol{z} - \mathbf{E}[\boldsymbol{z}]),$$

$$\mathrm{BN}(\tilde{\boldsymbol{z}}) = \begin{pmatrix} \gamma_{rr} & \gamma_{ri} \\ \gamma_{ri} & \gamma_{ii} \end{pmatrix} \tilde{\boldsymbol{z}} + \boldsymbol{\beta}. \tag{A.6}$$

The parameters $\gamma_{r(i)r(i)}$ and β are trainable. Each convolutional layer is composed of a convolutional operation, a CReLU and a BN layer. The first FC layer is composed of a FC operation and a CReLU. The last FC layer generates the final output directly via a FC operation. The operations defined above are differentiable, which means the neural network could be trained efficiently with back-propagation. The gradient is calculated with respect to the real-valued cost function $\mathcal{L}$ as

$$\nabla_{\mathcal{L}}(z) = \frac{\partial \mathcal{L}}{\partial z} = \frac{\partial \mathcal{L}}{\partial z_r} + \mathrm{i}\frac{\partial \mathcal{L}}{\partial z_i} = \Re(\nabla_{\mathcal{L}}(z)) + \mathrm{i}\Im(\nabla_{\mathcal{L}}(z)). \tag{A.7}$$

The back-propagation updates the complex-valued parameter $t = t_r + \mathrm{i}t_i$ of the neural network by

$$\nabla_{\mathcal{L}}(t) = \frac{\partial \mathcal{L}}{\partial t} = \frac{\partial \mathcal{L}}{\partial t_r} + \mathrm{i}\frac{\partial \mathcal{L}}{\partial t_i} \tag{A.8}$$

$$= \frac{\partial \mathcal{L}}{\partial z_r}\frac{\partial z_r}{\partial t_r} + \frac{\partial \mathcal{L}}{\partial z_i}\frac{\partial z_i}{\partial t_r} + \mathrm{i}\left(\frac{\partial \mathcal{L}}{\partial z_r}\frac{\partial z_r}{\partial t_i} + \frac{\partial \mathcal{L}}{\partial z_i}\frac{\partial z_i}{\partial t_i}\right) \tag{A.9}$$

$$= \frac{\partial \mathcal{L}}{\partial z_r}\left(\frac{\partial z_r}{\partial t_r} + \mathrm{i}\frac{\partial z_r}{\partial t_i}\right) + \frac{\partial \mathcal{L}}{\partial z_i}\left(\frac{\partial z_i}{\partial t_r} + \mathrm{i}\frac{\partial z_i}{\partial t_i}\right) \tag{A.10}$$

$$= \Re(\nabla_{\mathcal{L}}(z))\left(\frac{\partial z_r}{\partial t_r} + \mathrm{i}\frac{\partial z_r}{\partial t_i}\right) \tag{A.11}$$

$$+ \Im(\nabla_{\mathcal{L}}(z))\left(\frac{\partial z_i}{\partial t_r} + \mathrm{i}\frac{\partial z_i}{\partial t_i}\right), \tag{A.12}$$

which could be implemented using Pytorch [45].

Appendix B. Computing the GME of quantum pure states

We employ the algorithm proposed in [42] to compute the GME for an arbitrary quantum pure state. The algorithm is based on a tensor version of the Gauss–Seidel method for computing unitary eigenpairs (U-eigenpairs) of a non-symmetric complex tensor $\mathcal{A}$ which corresponds to the given quantum pure state.

Algorithm 1 [42]. Computing the U-eigenpairs of an $n_1 \times \cdots \times n_m$ non-symmetric complex tensor $\mathcal{A}$.

Step 1 (initial step): let $\mathcal{S} = \mathbf{sym}(\mathcal{A})$ be the symmetric embedding of $\mathcal{A}$, and $n = n_1 + \cdots + n_m$. Choose a starting point $\mathbf{x}_0 \in \mathbf{C}^n$ with $\|\mathbf{x}_0\| = 1$, and $0 < \alpha_{\mathcal{S}} \in \mathbf{R}$. Let $\lambda_0 = \mathcal{S}^*\mathbf{x}_0^m$.

Step 2 (iterating step):

for $k = 1, 2, \ldots,$ **do**

$$\hat{\mathbf{x}}_k = \lambda_{k-1}\mathcal{S}\mathbf{x}_{k-1}^{*m-1} + \alpha_{\mathcal{S}}\mathbf{x}_{k-1}, \tag{B.1}$$

$$\mathbf{x}_k = \hat{\mathbf{x}}_k/\|\hat{\mathbf{x}}_k\|, \tag{B.2}$$

$$\lambda_k = \mathcal{S}^*\mathbf{x}_k^m. \tag{B.3}$$

end for.
return:
Unitary symmetric eigenpair (US-pair): $\lambda_{\mathcal{S}} = |\lambda_k|$, and $\mathbf{x} = (\frac{\lambda_{\mathcal{S}}}{\lambda_k})^{1/m}\mathbf{x}_k$.
Let $\mathbf{x} = (\mathbf{x}^{(1)\top}, \ldots, \mathbf{x}^{(m)\top})^\top$, $\mathbf{x}^{(i)} \in \mathbf{C}^{n_i}$, for all $i = 1 : m$.
U-eigenvalue $\lambda_{\mathcal{A}} = \frac{(\sqrt{m})^m}{m!}\lambda_{\mathcal{S}}$.
U-eigenvector $\{\sqrt{m}\mathbf{x}^{(1)}, \ldots, \sqrt{m}\mathbf{x}^{(m)}\}$.

ORCID iDs

Yu Pan https://orcid.org/0000-0001-6900-4016

References

[1] Horodecki R, Horodecki P, Horodecki M and Horodecki K 2009 Quantum entanglement *Rev. Mod. Phys.* **81** 865
[2] Streltsov A, Adesso G and Plenio M B 2017 Colloquium: quantum coherence as a resource *Rev. Mod. Phys.* **89** 041003
[3] Chitambar E and Gour G 2019 Quantum resource theories *Rev. Mod. Phys.* **91** 025001
[4] Nielson M A and Chuang I I 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)
[5] Deutsch I H 2020 Harnessing the power of the second quantum revolution *PRX Quantum* **1** 020101
[6] Peres A 1996 Separability criterion for density matrices *Phys. Rev. Lett.* **77** 1413
[7] Gurvits L 2003 Classical deterministic complexity of Edmonds' problem and quantum entanglement *Proc. of the 35th Annual ACM Symp. on Theory of Computing* pp 10–9
[8] Horodecki R, Horodecki M and Horodecki P 1996 Teleportation, Bell's inequalities and inseparability *Phys. Lett.* A **222** 21–5
[9] Terhal B M 2000 Bell inequalities and the separability criterion *Phys. Lett.* A **271** 319–26
[10] Gühne O and Tóth G 2009 Entanglement detection *Phys. Rep.* **474** 1–75
[11] Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N and Lloyd S 2017 Quantum machine learning *Nature* **549** 195–202
[12] Xiao J, Yan Y, Zhang J and Tang Y 2010 A quantum-inspired genetic algorithm for *k*-means clustering *Expert Syst. Appl.* **37** 4966–73
[13] Lloyd S, Mohseni M and Rebentrost P 2014 Quantum principal component analysis *Nat. Phys.* **10** 631–3
[14] Tang E 2019 A quantum-inspired classical algorithm for recommendation systems *Proc. of the 51st Annual ACM SIGACT Symp. on Theory of Computing* pp 217–28
[15] Bukov M, Day A G R, Sels D, Weinberg P, Polkovnikov A and Mehta P 2018 Reinforcement learning in different phases of quantum control *Phys. Rev.* X **8** 031086
[16] Chapman R J, Ferrie C and Peruzzo A 2016 Experimental demonstration of self-guided quantum tomography *Phys. Rev. Lett.* **117** 040402
[17] Magesan E, Gambetta J M, Córcoles A D and Chow J M 2015 Machine learning for discriminating quantum measurement trajectories and improving readout *Phys. Rev. Lett.* **114** 200501
[18] Hentschel A and Sanders B C 2010 Machine learning for precise quantum measurement *Phys. Rev. Lett.* **104** 063603
[19] Carleo G and Troyer M 2017 Solving the quantum many-body problem with artificial neural networks *Science* **355** 602–6
[20] Huang L and Wang L 2017 Accelerated Monte Carlo simulations with restricted Boltzmann machines *Phys. Rev.* B **95** 035105
[21] Carrasquilla J and Melko R G 2017 Machine learning phases of matter *Nat. Phys.* **13** 431–4
[22] Lu S *et al* 2018 Separability-entanglement classifier via machine learning *Phys. Rev.* A **98** 012315
[23] Yang M *et al* 2019 Experimental simultaneous learning of multiple nonclassical correlations *Phys. Rev. Lett.* **123** 190401
[24] Ma Y-C and Yung M-H 2018 Transforming Bell's inequalities into state classifiers with machine learning *npj Quantum Inf.* **4** 34
[25] Liu N and Rebentrost P 2018 Quantum machine learning for quantum anomaly detection *Phys. Rev.* A **97** 042315
[26] Liang J-M, Shen S-Q, Li M and Li L 2019 Quantum anomaly detection with density estimation and multivariate Gaussian distribution *Phys. Rev.* A **99** 052310
[27] Bennett C H, Brassard G, Crépeau C, Jozsa R, Peres A and Wootters W K 1993 Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels *Phys. Rev. Lett.* **70** 1895
[28] Werner E E 1989 High-risk children in young adulthood: a longitudinal study from birth to 32 years *Am. J. Orthopsychiatry* **59** 72–81
[29] Dür W, Briegel H-J, Cirac J I and Zoller P 1999 Quantum repeaters based on entanglement purification *Phys. Rev.* A **59** 169
[30] Baldi P 2012 Autoencoders, unsupervised learning, and deep architectures *Proc. of ICML Workshop on Unsupervised and Transfer Learning* pp 37–49
[31] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in Neural Information Processing Systems* pp 2672–80
[32] Arjovsky M, Chintala S and Bottou L 2017 Wasserstein generative adversarial networks *Int. Conf. on Machine Learning* pp 214–23
[33] Chicco D 2021 Siamese neural networks: an overview *Artificial Neural Networks* (Springer) pp 73–94
[34] Koch G, Zemel R and Salakhutdinov R 2015 Siamese neural networks for one-shot image recognition *ICML Deep Learning Workshop* vol 2 (Lille)
[35] Tax D M J and Duin R P W 2004 Support vector data description *Mach. Learn.* **54** 45–66
[36] Hughes L H, Schmitt M, Mou L, Wang Y and Zhu X X 2018 Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN *IEEE Geosci. Remote Sens. Lett.* **15** 784–8

[37] Isola P, Zhu J-Y, Zhou T and Efros A A 2017 Image-to-image translation with conditional adversarial networks *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 1125–34
[38] Zyczkowski K and Sommers H-J 2001 Induced measures in the space of mixed quantum states *J. Phys. A: Math. Gen.* **34** 7111
[39] Brown C D and Davis H T 2006 Receiver operating characteristics curves and related decision measures: a tutorial *Chemometr. Intell. Lab. Syst.* **80** 24–38
[40] Zhou X-H, McClish D K and Obuchowski N A 2009 *Statistical Methods in Diagnostic Medicine* vol 569 (New York: Wiley)
[41] Acin A, Bruß D, Lewenstein M and Sanpera A 2001 Classification of mixed three-qubit states *Phys. Rev. Lett.* **87** 040401
[42] Zhang M, Ni G and Zhang G 2020 Iterative methods for computing U-eigenvalues of non-symmetric complex tensors with application in quantum entanglement *Comput. Optim. Appl.* **75** 779–98
[43] van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605
[44] Trabelsi C *et al* 2018 Deep complex networks *Int. Conf. on Learning Representations*
[45] Paszke A *et al* 2017 Automatic differentiation in Pytorch *NIPS 2017 Workshop on Autodiff* (Long Beach, California, USA) https://openreview.net/forum?id=BJJsrmfCZ

Witnessing Entanglement

Marisol N. Beck* and M. Beck†

*Department of Physics, Harvey Mudd College, 301 Platt Blvd., Claremont, CA 91711
† Department of Physics, Whitman College, 345 Boyer Ave., Walla Walla, WA 99362

Abstract: An entangled state of a two-particle system is a quantum state that cannot be separated—it cannot be written as the product of states of the individual particles. One way to tell if a system is entangled is to use it to violate a Bell inequality (such as the Clauser-Horne-Shimony-Holt, CHSH, inequality), because entanglement is necessary to violate these inequalities. However, there are other, more efficient measurements that determine whether or not a system is entangled; an operator that corresponds to such a measurement is referred to as an entanglement witness. We present the theory of witness operators, and an undergraduate experiment that measures an entanglement witness for the joint polarization state of two photons. We are able to produce states for which the expectation value of the witness operator is entangled by more than 160 standard deviations.

Keywords: Quantum Mechanics, Entanglement, Quantum Measurement, Quantum Information.
PACS: 03.67.Mn, 03.67.Bg, 42.50.Dv.

INTRODUCTION

Entanglement is a (perhaps *the*) feature that distinguishes quantum mechanics from classical mechanics. Entanglement is necessary for a diverse range of uniquely quantum mechanical effects such as quantum cryptography, quantum teleportation and quantum computing.[1]

Mathematically, entangled states are those quantum states that cannot be written as the product of the states of the individual particles. Thus, if $|\psi_{ent}\rangle$ represents an entangled state of a bipartite system, and $|\psi_A\rangle$ and $|\psi_B\rangle$ are the states of the individual particles, then

$$|\psi_{ent}\rangle \neq |\psi_A\rangle \otimes |\psi_B\rangle \ , \qquad (1)$$

where $\otimes$ represents the direct product.

In Eq. (1) $|\psi_{ent}\rangle$ is an entangled pure state. It has been shown that for every bipartite pure-state, there exists a Bell inequality that is violated;[2,3] this means that there exists, at least in principle, a method to experimentally detect that entanglement.

However, real experimental systems never exist in pure states. One must assume that the state of an experiment will yield a mixed state that must be described by density operator $\hat{\rho}$.[4] A mixed state is separable, and hence not entangled, if it can be written as a weighted sum of product states:

$$\hat{\rho}_{sep} = \sum_i \ p_i \hat{\rho}_{Ai} \otimes \hat{\rho}_{Bi} \ , \qquad (2)$$

where the p_i's are nonnegative real numbers, and the normalization condition is that they must sum to 1.

An observable that is able to detect entanglement is referred to as an entanglement witness.[5,6] Bell inequalities were the first entanglement witnesses, but there are other, more efficient, observables that are capable of detecting entanglement. For example, the minimum number of measurements needed to determine a Bell inequality for bipartite qubits (two, 2-state particles) is four, whereas it is possible to construct an entanglement witness for these same qubits that requires only three measurements.[7] The reason Bell inequalities require more measurements is because they are capable of ruling out any local-realistic model, whereas other entanglement witnesses assume the validity of quantum mechanics, and merely seek to determine whether or not a particular system is entangled.

Experiments with entangled photons have been previously performed in undergraduate laboratories.[4,8-12] These experiments include tests of Bell inequalities, which prove that the states used in those experiments were entangled. However, we know of no previous undergraduate experiments that measure the types of entanglement witnesses that we describe here. These witnesses require only three measurements, not four. Furthermore, we demonstrate that our witness operators are able to detect entanglement in situations where the Clauser-Horne-Shimony-Holt, CHSH, inequality,[8,9] which is the most commonly used Bell inequality, does not.

A full discussion of mixed-state density operators and witness operators is well beyond the scope of this article. For a discussion of density operators that is accessible to undergraduates, see Ref. [4]. For a more

2015 BFY Proceedings, edited by Eblen-Zayas, Behringer, and Kozminski; Peer-reviewed, doi:10.1119/bfy.2015.pr.003

complete discussion of witness operators, see Refs. [3] and [6].

THEORY

Schmidt Decomposition

Before discussing the general problem of identifying entanglement in arbitrary mixed state systems, let's first consider entanglement of pure states. Suppose that system *A* has dimension *N* and system *B* has dimension *M*. An arbitrary pure state of the joint system can be written as

$$\begin{aligned} |\psi\rangle &= \sum_{i}^{N} \sum_{j}^{M} c_{ij} |\alpha_i\rangle_A \otimes |\beta_j\rangle_B \\ &= \sum_{i}^{N} \sum_{j}^{M} c_{ij} |\alpha_i \beta_j\rangle . \end{aligned} \tag{3}$$

The Schmidt decomposition of $|\psi\rangle$ determines two new sets of basis vectors $|a_i\rangle_A$ and $|b_i\rangle_B$, such that

$$|\psi\rangle = \sum_{i}^{R} \lambda_i |a_i b_i\rangle \ . \tag{4}$$

The number *R* is called the Schmidt rank of the system, and $R \le \min(N, M)$. This is a simplification, because we have gone from a double sum to a single sum. The fact that the Schmidt decomposition of $|\psi\rangle$ exists is proven in Ref. [1].

The Schmidt decomposition is useful for several reasons. The Schmidt rank of any pure product state is 1; any pure state with $R > 1$ is entangled. We'll see another use for the Schmidt decomposition below.

Witness Operators

An observable $\hat{W}$ is an entanglement witness if

$$\langle \hat{W} \rangle = \mathrm{Tr}\left(\hat{W} \hat{\rho}_{sep}\right) \ge 0 \tag{5}$$

for all separable states $\hat{\rho}_{sep}$, and

$$\langle \hat{W} \rangle = \mathrm{Tr}\left(\hat{W} \hat{\rho}_{ent}\right) < 0 \tag{6}$$

for at least one entangled state $\hat{\rho}_{ent}$.[3,5,6] Here Tr() refers to the trace of an operator. This means that if one measures $\langle \hat{W} \rangle < 0$, one knows that the state $\hat{\rho}$ is entangled.

There are different ways to construct witness operators. The technique that we use is to note that if our experimentally produced state is "close enough" (in Hilbert space) to a particular entangled pure state $|\psi_{ent}\rangle$, it will be entangled as well. As such we construct the witness operator[6]

$$\hat{W} = \alpha \hat{1} - \hat{\rho}_{ent} = \alpha \hat{1} - |\psi_{ent}\rangle\langle\psi_{ent}| \ . \tag{7}$$

In order to ensure that this operator meets the definition of an entanglement witness, the constant α is chosen to have the minimum value possible such that $\hat{W}$ satisfies Eq. (5):

$$\langle \hat{W} \rangle = \alpha \langle \hat{1} \rangle - \mathrm{Tr}\left(|\psi_{ent}\rangle\langle\psi_{ent}| \hat{\rho}_{sep}\right) \ge 0 \ . \tag{8}$$

We thus require α to be given by

$$\alpha = \max \mathrm{Tr}\left(|\psi_{ent}\rangle\langle\psi_{ent}| \hat{\rho}_{sep}\right). \tag{9}$$

Actually performing the maximization in Eq. (9) is beyond the scope of this article. It can be shown that α is given by the square of the maximum Schmidt coefficient of $|\psi_{ent}\rangle$, $\lambda_{i\max}$.[6,13]

The two entangled states we are interested in detecting are the Bell states of two photons

$$|\Phi^{\pm}\rangle = \frac{1}{\sqrt{2}}\left(|HH\rangle \pm |VV\rangle\right) , \tag{10}$$

where *H* and *V* correspond to horizontally and vertically polarized photons. This is the Schmidt decomposition of $|\Phi^{\pm}\rangle$, so the maximum Schmidt coefficient is $1/\sqrt{2}$, and the witness operators are

$$\begin{aligned} \hat{W}^{\pm} &= \frac{1}{2}\hat{1} - |\Phi^{\pm}\rangle\langle\Phi^{\pm}| \\ &= \frac{1}{2}\Big[\hat{1} - |HH\rangle\langle HH| - |VV\rangle\langle VV| \\ &\quad \mp \left(|HH\rangle\langle VV| + |VV\rangle\langle HH|\right)\Big]. \end{aligned} \tag{11}$$

In the laboratory, we are able to perform local, projective measurements. That is, both Alice and Bob perform projective measurements on their respective particles. The first two terms after the $\hat{1}$ in Eq. (11) take this form, but the two terms in parentheses don't.

However, we recognize that Alice and Bob are not limited to performing measurements in the horizontal-vertical basis. Define the diagonal and antidiagonal ($\pm 45^{o}$ linear), and the left- and right- circular polarization states as

$$|D\rangle = \frac{1}{\sqrt{2}}\left(|H\rangle + |V\rangle\right) \ , \ |A\rangle = \frac{1}{\sqrt{2}}\left(|H\rangle - |V\rangle\right) \quad (12)$$

$$|L\rangle = \frac{1}{\sqrt{2}}\left(|H\rangle + i|V\rangle\right) \ , \ |R\rangle = \frac{1}{\sqrt{2}}\left(|H\rangle - i|V\rangle\right). \quad (13)$$

We rewrite our witness using these operators as:

$$\hat{W}^{\pm} = \frac{1}{2}\Big[\hat{1} - |HH\rangle\langle HH| - |VV\rangle\langle VV| \mp \left(|DD\rangle\langle DD| + |AA\rangle\langle AA| - |LL\rangle\langle LL| - |RR\rangle\langle RR|\right)\Big]. \quad (14)$$

Defining $P(a,b)$ to be the joint probability that Alice measures her photon to have polarization a, and Bob measures his photon to have polarization b, we find that the expectation value of the witness operators is

$$\langle \hat{W}^{\pm} \rangle = \frac{1}{2}\Big\{1 - P(H,H) - P(V,V) \mp \big[P(D,D) + P(A,A) - P(L,L) - P(R,R)\big]\Big\}. \quad (15)$$

EXPERIMENTS

Our experiments are similar to those performed in Ref. [7], but we use equipment that is currently found in many undergraduate laboratories.[4] The experimental apparatus is shown in Fig. 1. A 405 nm laser diode pumps a pair of Type-I beta-barium borate crystals, whose axes are oriented at right angles with respect to each other. Down converted photons pass through a series of wave plates and polarizing beam splitters, before being focused onto multimode optical fibers and detected with single-photon counting modules.

The polarization states of the down converted photon pairs are adjusted using the techniques of previous experiments.[4,8] The states that we are trying to produce take the form

$$|\Phi(\phi)\rangle = \frac{1}{\sqrt{2}}\left(|HH\rangle + e^{i\phi}|VV\rangle\right). \quad (16)$$

The birefringent plate in the pump beam is mounted on a tilt stage with a micrometer, and is used to adjust the relative phase ϕ; note that $\phi = 0$ yields $|\Phi^{+}\rangle$ and $\phi = \pi$ yields $|\Phi^{-}\rangle$. The techniques described in Refs. 4 and 8 allow us to easily determine $\phi = 0$ and $\phi = \pi$. We extrapolate between these two tilt angles to determine the phase angle of the state.

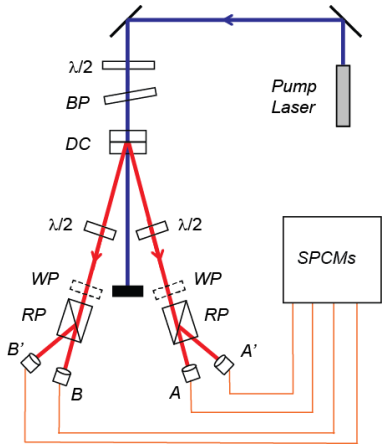


FIGURE 1. The experimental apparatus. Here λ/2 denotes a half-wave plate, *BP* denotes a birefringent plate, *DC* denotes down conversion crystals, *WP* denotes the wave plates used to do the measurement projections, *RP* denotes a Rochon polarizer, and *SPCMs* are the single-photon counting modules.

However, our experimentally produced states are not pure. We model our states as

$$\hat{\rho} = p|\Phi(\phi)\rangle\langle\Phi(\phi)| + \frac{1-p}{2}\left(|HH\rangle\langle HH| + |VV\rangle\langle VV|\right). \quad (17)$$

This density operator represents our photons as being in the entangled state $|\Phi(\phi)\rangle$ with probability p, and in an equal mixture of the states $|HH\rangle$ and $|VV\rangle$ with probability $1-p$.

With the optional wave plates removed (see Fig. 1) horizontally polarized photon pairs are directed to detectors *A* and *B*, and vertically polarized photons are directed to detectors *A'* and *B'*. We can thus measure the probabilities $P(H,H)$ and $P(V,V)$. The probabilities of detecting diagonal and antidiagonal photon pairs are obtained by inserting properly oriented half-wave plates before the Rochon polarizers. To measure the circular polarization probabilities we insert properly oriented quarter-wave plates.

Figure 2 shows the experimental data for our two witnesses, and the CHSH parameter *S*.[8,9] In Fig. 2(a) we see that when we are creating $|\Phi^{+}\rangle$ (ϕ near 0), $\langle \hat{W}^{+} \rangle$ indicates that the state is entangled, and $\langle \hat{W}^{-} \rangle$ does not. This is as we would expect, because $\hat{W}^{+}$ is

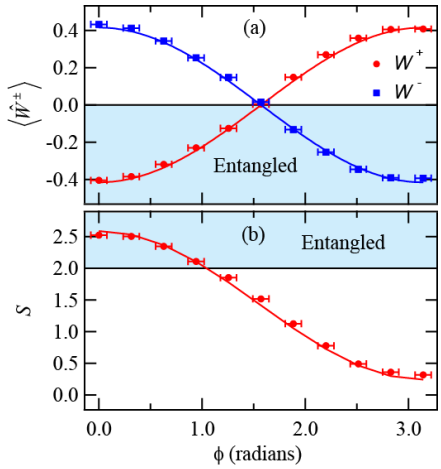


FIGURE 2. (a) $\langle \hat{W}^{+} \rangle$ (red circles) and $\langle \hat{W}^{-} \rangle$ (blue squares) are plotted as a function of the entangled state phase, ϕ. (b) The CHSH parameter S is plotted as a function of the entangled state phase, ϕ. The points are experimental data, while the solid lines are theoretical predictions. Statistical (vertical) error bars are smaller than the markers. Horizontal error bars are $\pm\pi/40$, which is our best estimate of how accurately we can set the phases.

constructed to witness this entangled state, while $\hat{W}^{-}$ is not. Their behavior switches as ϕ approaches π and we are constructing state $|\Phi^{-}\rangle$. Note that the version of the CHSH inequality that we use detects entanglement in $|\Phi^{+}\rangle$ when $S > 2$. However, $\hat{W}^{+}$ does a "better" job of detecting this entanglement: $\langle \hat{W}^{+} \rangle$ indicates that the point at $\phi \cong 1.25$ rad is entangled, while S does not.

The expectation values $\langle \hat{W}^{\pm} \rangle$ are obtained from the same data, but computed differently. The data for S is obtained separately because it requires different measurement settings. Our technique for obtaining the measurements in Fig. 2 is to set the value of ϕ, measure $\langle \hat{W}^{\pm} \rangle$ and S one after the other, then change ϕ and repeat. We note that at $\phi = 0$ in Fig. 2 $\langle \hat{W}^{+} \rangle = -0.4042 \pm 0.0025$, which indicates that the state is entangled by over 160 standard deviations for 300 s of counting time. This same state yields $S = 2.521 \pm 0.012$, which violates the classical inequality by 40 standard deviations for 400 s of counting time.

For states described by Eq. (17), the theoretical expectation values of the witness operators, are

$$\langle \hat{W}^{\pm} \rangle = \mp \frac{p}{2} \cos\phi \ . \qquad (18)$$

We treat p as a free parameter, and use it to fit our data for $\langle \hat{W}^{+} \rangle$; we find that $p = 0.83 \pm 0.01$. Once this value has been determined for $\langle \hat{W}^{+} \rangle$, we use it to determine the theoretical predictions for $\langle \hat{W}^{-} \rangle$ and S. Thus, we use one free parameter for all three theoretical curves shown in Fig. 2.

CONCLUSIONS

We have experimentally measured the expectation values of two different entanglement witness operators $\langle \hat{W}^{\pm} \rangle$ in an undergraduate laboratory, and compared them to measurements of the CHSH parameter S. Determining $\langle \hat{W}^{\pm} \rangle$ is "easier" in that they require only three measurements, as compared to four measurements for S. The witness operators also indicate entanglement for states that S does not, and they yield a larger violation of classical physics (in terms of the number of standard deviations that a classical inequality is violated).

ACKNOWLEDGMENTS

We acknowledge the support of Whitman College.

REFERENCES

1. M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, 2000).
2. N. Gisin and A. Peres, Phys. Lett. A **162**, 15-17 (1992).
3. B. M. Terhal, Theor. Comput. Sci. **287**, 313-335 (2002).
4. M. Beck, *Quantum Mechanics: Theory and Experiment* (Oxford Univ. Press, Oxford, 2012).
5. Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki, Phys. Lett. A **223**, 1-8 (1996).
6. O. Gühne and G. Toth, Phys. Rep. **474**, 1-75 (2009).
7. M. Barbieri et al., Phys. Rev. Lett. **91**, 4 (2003).
8. D. Dehlinger and M. W. Mitchell, Am. J. Phys. **70**, 898-902 (2002).
9. D. Dehlinger and M. W. Mitchell, Am. J. Phys. **70**, 903-910 (2002).
10. J. A. Carlson, M. D. Olmstead, and M. Beck, Am. J. Phys. **74**, 180-186 (2006).
11. E. J. Galvez, Am. J. Phys. **78**, 510-519 (2010).
12. E. Dederick and M. Beck, Am. J. Phys. **82**, 962-971 (2014).
13. Mohamed Bourennane et al., Phys. Rev. Lett. **92**, 087902 (2004).

entropy

MDPI

Article

Spatial Entanglement of Fermions in One-Dimensional Quantum Dots

Ivan P. Christov [1,2]

[1] Physics Department, Sofia University, 1164 Sofia, Bulgaria; ivan.christov@phys.uni-sofia.bg
[2] Institute of Electronics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria

Abstract: The time-dependent quantum Monte Carlo method for fermions is introduced and applied in the calculation of the entanglement of electrons in one-dimensional quantum dots with several spin-polarized and spin-compensated electron configurations. The rich statistics of wave functions provided by this method allow one to build reduced density matrices for each electron, and to quantify the spatial entanglement using measures such as quantum entropy by treating the electrons as identical or distinguishable particles. Our results indicate that the spatial entanglement in parallel-spin configurations is rather small, and is determined mostly by the spatial quantum nonlocality introduced by the ground state. By contrast, in the spin-compensated case, the outermost opposite-spin electrons interact like bosons, which prevails their entanglement, while the inner-shell electrons remain largely at their Hartree–Fock geometry. Our findings are in close correspondence with the numerically exact results, wherever such comparison is possible.

Keywords: quantum correlations; quantum entanglement; quantum Monte Carlo method

Citation: Christov, I.P. Spatial Entanglement of Fermions in One-Dimensional Quantum Dots. *Entropy* **2021**, *23*, 868. https://doi.org/10.3390/e23070868

Academic Editor: Fabio Benatti

Received: 10 June 2021
Accepted: 5 July 2021
Published: 7 July 2021

1. Introduction

During the past few decades there has been an increasing interest in developing new models and computational tools to address the fundamental and practical challenges related to quantum correlations and entanglement, in connection with their potential application in newly emerging quantum technologies [1]. The properties of composite systems of quantum particles are expected to play an important role in information processing, as well as in devices for manipulating systems of atoms and molecules. While various algebraic operator methods have been used to characterize entanglement in spin systems [2–4], an efficient approach to assess the spatial entanglement in many-body quantum systems together with its evolution over time is still lacking. It is well known that the correlated non-relativistic particle motion described by the time-dependent Schrödinger equation (SE) is tractable for only a limited number of cases. While solvable numerically for few particles in 1D and 2D, the direct numerical solution of the SE scales exponentially with system size, and is therefore beyond the capabilities of today's computers. That exponential time scaling is usually attributed to the nonlocal quantum effects that result from the dependence of the wave function $\Psi(\mathbf{r}_1, \ldots, \mathbf{r}_N, t)$ on the coordinates of all interacting particles. The standard approaches to ameliorate the workload are to reduce the many-body SE to a set of coupled single-body equations, where the most prominent are the mean-field approaches: the Hartree–Fock (HF) method [5] and density-functional theory (DFT) [6]. That reduction, however, occurs at the price of neglecting the detailed fluctuating forces between the electrons and replacing them with averages, thus totally ignoring the dynamic quantum correlations in the HF method, while DFT reduces the many-body problem to a single-body problem of non-interacting electrons moving in an effective exchange-correlation potential, which is generally unknown and suffers self-interaction issues. More accurate but more computationally expensive are the multi-configuration time-dependent Hartree–Fock method [7], and the full configuration interaction method [8]. Another approach that has gained much attention lately is the density matrix renormalization group method [9],

which allows one to treat correlated 1D many-body problems with good accuracy; however, its application to higher dimensions and arbitrary potentials has been challenging thus far.

A different class of methods to tackle the quantum many-body problem includes the quantum Monte Carlo methods [10], which allow one to accurately calculate the electronic structures of atoms, molecules, nanostructures, and condensed-matter systems at a fully correlated level. For example, the diffusion quantum Monte Carlo (DMC) method uses random particles (walkers) whose evolution towards the ground state of the system involves a combination of diffusion and branching—which, however, prevent its use for real-time-dependent processes where causality is of primary importance. Moreover, the artificial nature of the many-body wave function in configuration space used in the DMC method prevents the calculation of some important quantities other than the energy. The recently introduced time-dependent quantum Monte Carlo (TDQMC) method [11–13] employs concurrent ensembles of walkers and wave functions for each electron, where each wave function is associated with a separate walker (particle–wave dichotomy), and these evolve in physical spacetime where no initial guess for the many-body wave function is needed. Recently, we have applied the TDQMC method to analyze the ground state preparation for simple bosonic systems with several particles in one and two dimensions, with a good tradeoff between scaling and accuracy [14]. In this work we apply the TDQMC method to several interacting fermions in one-dimensional quantum dots, where the obtained reduced density matrices allow us to quantify the entanglement between the electrons considered to be identical or distinguishable particles. We consider some proof-of-principle aspects of the TDQMC method for fermions, rather than practical matters concerning quantum dots. Entanglement in two-electron quantum dots, atoms, and molecules has been studied elsewhere [15–18].

2. Methods

The TDQMC method transforms the standard Hartree–Fock (HF) equations into a set of coupled stochastic equations capable of describing the correlated particle motion. That transformation is based on the physical assumption that the modulus square of the single-body wave function in coordinate (physical) space may be thought of as an envelope (or kernel density estimation) of the distribution of a finite number of particles (walkers). In this way, for each electron in an atom a large set of single-body wave functions that reside in physical spacetime is created, where each wave function responds to the multicore potential due to both the nucleus and the walkers of the rest of the electrons. The crucial point in this picture is that it allows for each walker for a given electron to interact with the walkers of any other electron through weighted Coulomb potential, thus naturally incorporating the spatial quantum nonlocality. Then, from the evolution of the walker distributions one can evaluate quantum observables without resorting to the many-body wave function. Formally, we start from the HF equation for the i$^{\text{th}}$ from a total of N electrons, within a single-determinant ansatz [19]:

$$i\hbar\frac{\partial}{\partial t}\varphi_i(\mathbf{r}_i,t)=\left[-\frac{\hbar^2}{2m}\nabla_i^2+V_{en}(\mathbf{r}_i)+V_{ee}^{HF}(\mathbf{r}_i,t)\right]\varphi_i(\mathbf{r}_i,t) \tag{1}$$

where $V_{en}(\mathbf{r}_i)$ is the electron–nuclear potential, and the HF electron–electron potential reads:

$$V_{ee}^{HF}(\mathbf{r}_i,t)=V_{ee}^{H}(\mathbf{r}_i,t)+V_{ee}^{X}(\mathbf{r}_i,t) \tag{2}$$

where:

$$V_{ee}^{H}(\mathbf{r}_i,t)=\sum_{j\neq i}^{N}\int d\mathbf{r}_j V_{ee}(\mathbf{r}_i-\mathbf{r}_j)\left|\varphi_j(\mathbf{r}_j,t)\right|^2 \tag{3}$$

is the Hartree potential, and $V_{ee}^{X}(\mathbf{r}_i,t)$ is the exchange potential:

$$V_{ee}^{X}(\mathbf{r}_i,t) = -\sum_{j\neq i}^{N}\delta_{s_i,s_j}\int d\mathbf{r}_j V_{ee}(\mathbf{r}_i-\mathbf{r}_j)\varphi_i(\mathbf{r}_j,t)\varphi_j^{*}(\mathbf{r}_j,t)\varphi_j(\mathbf{r}_i,t)/\varphi_i(\mathbf{r}_i,t) \tag{4}$$

where $\varphi_i(\mathbf{r},t)$ satisfy the orthonormality property $\int \varphi_i(\mathbf{r},t)\varphi_j^{*}(\mathbf{r},t)d\mathbf{r} = \delta_{i,j}$, and the indices s_i, s_j denote the spins of the corresponding electrons. The inequality $j \neq i$ in the sums of Equations (3) and (4) stresses the fact that even though the self-interaction between the electrons is naturally canceled in the HF approximation, it is also not present in the Hartree approximation, where there is no exchange potential [19]. It is known that the wave functions $\varphi_i(\mathbf{r},t)$ of Equation (1) variationally minimize the system energy:

$$E^{HF} = \sum_{i=1}^{N}\left[-\frac{\hbar^2}{2m}\int \varphi_i^{*}(\mathbf{r}_i,\tau)\nabla_i^2\varphi_i(\mathbf{r}_i,\tau)d\mathbf{r}_i + \int V_{en}(\mathbf{r}_i)|\varphi_i(\mathbf{r}_i,\tau)|^2 d\mathbf{r}_i\right] + E_{ee}^{H} + E_{ee}^{X} \tag{5}$$

where the Hartree and exchange energies read:

$$E_{ee}^{H} = 0.5\sum_{i\neq j}^{N}\iint V_{ee}(\mathbf{r}_i-\mathbf{r}_j)|\varphi_i(\mathbf{r}_i,\tau)|^2\left|\varphi_j(\mathbf{r}_j,\tau)\right|^2 d\mathbf{r}_i d\mathbf{r}_j \tag{6}$$

and

$$E_{ee}^{X} = -0.5\sum_{i\neq j}^{N}\delta_{s_i,s_j}\iint V_{ee}(\mathbf{r}_i-\mathbf{r}_j)\varphi_i(\mathbf{r}_j,\tau)\varphi_j^{*}(\mathbf{r}_j,\tau)\varphi_j(\mathbf{r}_i,\tau)\varphi_i^{*}(\mathbf{r}_i,\tau)d\mathbf{r}_i d\mathbf{r}_j \tag{7}$$

respectively.

It is known that the Hartree–Fock approximation does not account for the dynamic electron–electron correlations beyond those due to the exchange interaction. In order to correct for this in the TDQMC methodology, we replace the HF wave function for each electron $\varphi_i(\mathbf{r},t)$ with a family of slightly different wave functions $\varphi_i(\mathbf{r},t) \rightarrow \varphi_i^{k}(\mathbf{r},t)$; k = 1, ... ,M [11–13], which allows us to further lower the system energy below the HF level. This is accomplished by applying a stochastic windowing to the distribution $\left|\varphi_j(\mathbf{r}_j,t)\right|^2$ in the Hartree potential $V_{ee}^{H}(\mathbf{r}_i,t)$ of Equation (3), by using a "window" function $K\left[\mathbf{r}_j,\mathbf{r}_j^{k}(t),\sigma_{j,i}\right]$ centered at a certain walker's trajectory $\mathbf{r}_j^{k}(t)$, which samples the distribution given by $\left|\varphi_j(\mathbf{r}_j,t)\right|^2$. The parameters $\sigma_{j,i}$ determine the widths of those "windows", such that the product $\left|\varphi_j(\mathbf{r}_j,t)\right|^2 K\left[\mathbf{r}_j,\mathbf{r}_j^{k}(t),\sigma_{j,i}\right]$ is different for each separate trajectory $\mathbf{r}_j^{k}(t)$. In this way, for each electron, Equations (1)–(4) are transformed into a set of M Hartree–Fock-like equations for the different replicas $\varphi_i^{k}(\mathbf{r}_i,t)$ of the initial HF wave function $\varphi_i(\mathbf{r}_i,t)$, each one attached to a separate trajectory $\mathbf{r}_j^{k}(t)$ (particle–wave dichotomy [13]):

$$\begin{aligned} i\hbar\tfrac{\partial}{\partial t}\varphi_i^{k}(\mathbf{r}_i,t) = \Big[&-\tfrac{\hbar^2}{2m}\nabla_i^2 + V_{en}(\mathbf{r}_i) + V_{eff}^{k}(\mathbf{r}_i,t) \\ &-\sum_{j\neq i}^{N}\delta_{s_i,s_j}\int d\mathbf{r}_j V_{ee}(\mathbf{r}_i-\mathbf{r}_j)\varphi_i^{k}(\mathbf{r}_j,t)\varphi_j^{k*}(\mathbf{r}_j,t)\varphi_j^{k}(\mathbf{r}_i,t)/\varphi_i^{k}(\mathbf{r}_i,t)\Big]\varphi_i^{k}(\mathbf{r}_i,t) \end{aligned} \tag{8}$$

where $\int \varphi_i^{k}(\mathbf{r},t)\varphi_j^{k*}(\mathbf{r},t)d\mathbf{r} = \delta_{i,j}$; i = 1, ... ,N, k = 1, ... ,M, and where:

$$V_{eff}^{k}(\mathbf{r}_i,t) = \sum_{j\neq i}^{N}\frac{1}{Z_{j,i}^{k}}\sum_{l}^{M}V_{ee}\left[\mathbf{r}_i,\mathbf{r}_j^{l}(t)\right]K\left[\mathbf{r}_j^{l}(t),\mathbf{r}_j^{k}(t),\sigma_{j,i}\right] \tag{9}$$

is the effective electron–electron interaction potential represented as a Monte Carlo (MC) convolution that incorporates the spatial quantum nonlocality by allowing each walker for a given electron to interact with a group of walkers of any other electron. In fact, it

can be seen from Equation (9) that the effective potential "seen" by the kth wave function for the ith electron involves interactions with a number of walkers that belong to the jth electron and lie within the nonlocal length $\sigma_{j,i}$ around $\mathbf{r}_j(t)$ [12–14]. For the Gaussian kernel we have:

$$K\left[\mathbf{r}_j, \mathbf{r}_j^k(t), \sigma_{j,i}\right] = \exp\left(-\frac{\left|\mathbf{r}_j - \mathbf{r}_j^k(t)\right|^2}{2\sigma_{j,i}^2}\right) \tag{10}$$

which determines the weighting factor in Equation (9) to be:

$$Z_{j,i}^k = \sum_{l=1}^{M} K\left[\mathbf{r}_j^l(t), \mathbf{r}_j^k(t), \sigma_{j,i}\right] \tag{11}$$

As seen from Equations (10) and (11), the limit $\sigma_{j,i} \to \infty$ where $K\left[\mathbf{r}_j, \mathbf{r}_j^k(t), \sigma_{j,i}\right] \to 1$ recovers the Hartree–Fock approximation—as opposed to the local interaction, where $\sigma_{j,i} \to 0$ and $K\left[\mathbf{r}_j, \mathbf{r}_j^k(t), \sigma_{j,i}\right] \to \delta\left(\mathbf{r}_j - \mathbf{r}_j^k(t)\right)$. It is clear, therefore, that $\sigma_{j,i}$ may serve as variational parameters to minimize the system energy between these two limiting cases.

The connection between the trajectories $\mathbf{r}_i^k(t)$ and the wave functions $\varphi_i^k(\mathbf{r}_i, t)$ is given by the walker's velocities (de Broglie–Bohm equation, e.g., in [20]):

$$\mathbf{v}_i^k(t) == \frac{\hbar}{m}\mathrm{Im}\left[\frac{\nabla_i \varphi_i^k(\mathbf{r}_i, t)}{\varphi_i^k(\mathbf{r}_i, t)}\right]_{\mathbf{r}_i = \mathbf{r}_i^k(t)} \tag{12}$$

for real-time propagation, and:

$$d\mathbf{r}_i^k(\tau) = \mathbf{v}^{Dk}{}_i d\tau + \boldsymbol{\eta}_i(\tau)\sqrt{\frac{\hbar}{m}d\tau} \tag{13}$$

for imaginary-time propagation, where the drift velocity reads:

$$\mathbf{v}^{Dk}{}_i(\tau) = \frac{\hbar}{m}\left[\frac{\nabla_i \varphi_i^k(\mathbf{r}_i, \tau)}{\varphi_i^k(\mathbf{r}_i, \tau)}\right]_{\mathbf{r}_i = \mathbf{r}_i^k(\tau)} \tag{14}$$

and $\boldsymbol{\eta}(\tau)$ is a Markovian stochastic process (see also the appendix in [21]). The striking similarity between the drift velocity of Equation (14) and the de Broglie–Bohm Equation (12) comes from the fact that both equations describe drift-diffusion processes in imaginary and in real time, respectively. It is seen that although the individual wave functions guide the corresponding walkers through Equation (12), the TDQMC method solves coupled single-body Hartree-Fock-like equations (e.g., Equation (8)) instead of using quantum potentials as in Bohmian mechanics [20].

Following the particle–wave dichotomy described above, the system energy can be calculated conveniently using both particle trajectories and wave functions:

$$\begin{aligned} E = \frac{1}{M}\sum_{k=1}^{M}\left[\sum_{i=1}^{N}\left[-\frac{\hbar^2}{2m}\frac{\nabla_i^2 \varphi_i^k(\mathbf{r}_i^k)}{\varphi_i^k(\mathbf{r}_i^k)} + V_{en}(\mathbf{r}_i^k)\right] + \sum_{i>j}^{N} V_{ee}(\mathbf{r}_i^k - \mathbf{r}_j^k)\right] & \begin{matrix}\mathbf{r}_i^k = \mathbf{r}_i^k(\tau) \\ \mathbf{r}_j^k = \mathbf{r}_j^k(\tau)\end{matrix} \\ -\frac{0.5}{M}\sum_{k=1}^{M}\sum_{i\neq j}^{N}\delta_{s_i,s_j}\int\int d\mathbf{r}_i d\mathbf{r}_j V_{ee}(\mathbf{r}_i - \mathbf{r}_j)\varphi_i^k(\mathbf{r}_j, t)\varphi_j^{k*}(\mathbf{r}_j, t)\varphi_j^k(\mathbf{r}_i, t)\varphi_i^{k*}(\mathbf{r}_i, t). & \end{aligned} \tag{15}$$

During the preparation of the ground state of the quantum system, the initial Monte Carlo ensembles of walkers and wave functions propagate in imaginary time (τ) toward a steady state, in accordance with Equations (8)–(14), for different values of the nonlocality parameters $\sigma_{j,i}$, until the energy of Equation (15) displays a minimum. Another alternative

to account for the exchange effects is to apply a short-range screening to the interaction potential in order to modify the repulsion between the same-spin electrons [22].

Considering the ensemble of waves $\varphi_i^k(\mathbf{r}_i, t)$ delivered by the TDQMC as random variables, one can build a reduced density matrix for the ith electron, which may serve as the variance–covariance matrix in the Hilbert space that carries important statistical information [13,23]:

$$\rho_i\left(\mathbf{r}_i, \mathbf{r}'_i, t\right) = \frac{1}{M}\sum_{k=1}^{M} \varphi_i^{k*}(\mathbf{r}_i, t)\varphi_i^k(\mathbf{r}'_i, t) \tag{16}$$

For example, the density matrix of Equation (16) allows one to easily calculate the spatial entanglement of a given electron state, which may serve also as a good measure for the overall accuracy of the calculation (e.g., [24]). Without entering the ongoing debate on entanglement witnesses, for opposite-spin electrons—where there are no exchange terms in HF and TDQMC equations—we employ the linear quantum entropy for distinguishable (non-identical) particles as a conventional measure for the spatial entanglement [25]:

$$S^i_{L\uparrow\downarrow}(t) = 1 - Tr\left(\rho_i^2\right) = 1 - \int \rho_i^2(\mathbf{r}_i, \mathbf{r}_i, t)d\mathbf{r}_i \tag{17}$$

while for N same-spin electrons the component of the entanglement that reflects the trivial minimum correlation due to the anti-symmetrization of the wave function can be eliminated [26–28], yielding:

$$S^i_{L\uparrow\uparrow}(t) = 1 - NTr\left(\rho_i^2\right) = 1 - N\int \rho_i^2(\mathbf{r}_i, \mathbf{r}_i, t)d\mathbf{r}_i \tag{18}$$

This definition ensures that for the wave functions used in HF approximation (or in general for any Slater rank 1 many-body state [29,30]) the linear entropy should vanish.

For indistinguishable (identical) particles, the spatial part of the 2N-body wave function can be represented in the simplest case as a product of normalized spin-up and spin-down Slater determinants [10]:

$$\Psi_i(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{2N}, t) = D_i^{\uparrow}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N, t)D_i^{\downarrow}(\mathbf{r}_{N+1}, \mathbf{r}_{N+2}, \ldots, \mathbf{r}_{2N}, t) \tag{19}$$

Thus, the entanglement between—for example—spin-up only electrons can be estimated using the reduced density matrix, averaged over the configurations provided by the TDQMC algorithm:

$$\rho_i^{\uparrow}\left(\mathbf{r}, \mathbf{r}', t\right) = \frac{1}{M}\sum_{k}^{M}\int D_i^{k\uparrow}(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N, t)D_i^{k\uparrow *}(\mathbf{r}', \mathbf{r}_2, \ldots, \mathbf{r}_N, t)d\mathbf{r}_2 \ldots d\mathbf{r}_N \tag{20}$$

where $D_i^{k\uparrow}$ are spin-up Slater determinants composed by the individual wave functions.

3. Results

As an example here we calculate the ground state of a quantum dot with parabolic core potential $V_{en}(\mathbf{r}_i) = \omega^2 r_i^2/2$ and with soft-core electron–electron Coulomb repulsion [31]:

$$V_{ee}\left[\mathbf{r}_i, \mathbf{r}_j\right] = \frac{e^2}{\sqrt{r^2 + a^2}} \tag{21}$$

where $r \equiv \left|\mathbf{r}_i - \mathbf{r}_j\right|$.

Within the formalism of Section 2 the degree of spatial correlation and, hence, the spatial entanglement, is controlled in TDQMC by the quantum nonlocal length $\sigma_{j,i}$ where, for bound electrons, higher $\sigma_{j,i}$ lead to lower correlation (entanglement) between the ith and the jth electron, and vice versa. Since the spatial extent of the electron cloud for the jth

electron is determined by the standard deviation s_j of the corresponding MC ensemble, the nonlocal length $\sigma_{j,i}$ is expected to be close to s_j:

$$\sigma_{j,i} = \alpha_{j,i}.s_j;\ \mathrm{j,i} = 1,\ldots,N, \tag{22}$$

where $\alpha_{j,i}$ may now serve as the variational parameters to minimizing the energy.

Since for parallel-spin electrons the eigenstates are orthogonal to one another, their overlap is small and, hence, the dynamic correlation between such states is expected to be smaller compared to the correlation between opposite-spin electrons. Here, we consider 1D quantum dots in two basic configurations: one is a spin-polarized configuration where each energy level is filled with just one same-spin electron, as seen in Figure 1a, and the other is a spin-compensated configuration where each level is occupied by two opposite-spin electrons (Figure 1b).

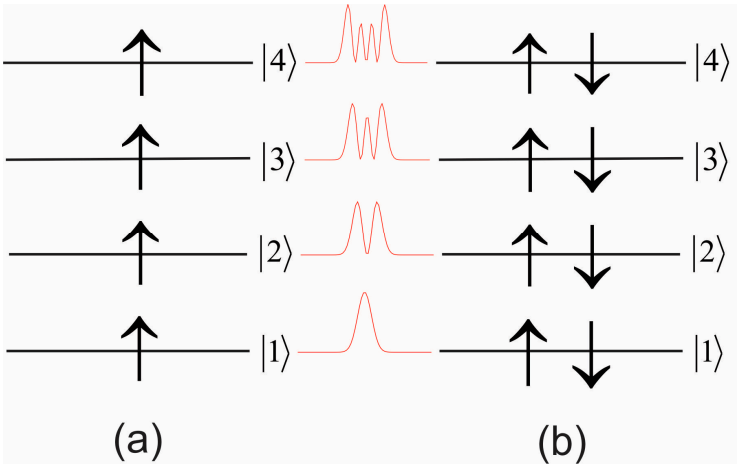


Figure 1. Energy level diagrams for spin-polarized (**a**) and spin-compensated (**b**) electrons in 1D quantum dots. The moduli square of the corresponding spatial orbitals are drawn with red.

We start with the ground state and the three excited states of a total of four same-spin electron configurations (Figure 1a), where due to the orthogonality of the spatial wave functions the dynamic correlations between the same-spin electrons are expected to be rather small. The system of coupled non-linear TDQMC equations (Equation (8)) is solved iteratively using both spilt-step Fourier and Crank–Nicolson numerical schemes, which give close results. The final states obtained are statistical variants of the corresponding Hartree–Fock wave functions for the different energy levels, with certain distortions due to the different effective interaction potentials $V^k_{eff}(\mathbf{r}_i, t)$ in Equation (8). Starting from preliminary calculated Hartree–Fock wave functions, after 200 steps of imaginary-time propagation of Equations (8)–(11) and Equations (13) and (14), for $\omega = 1$, and for different values of the variational parameter $\alpha_{j,i}$ of Equation (22), we find the energy minima for one, two, three, and four occupied levels in succession. Figure 2a shows these energies (green line), which are in a very good correspondence with the numerically exact energies (blue line) obtained from the direct numerical solution of the Schrödinger equation for up to four electrons in one spatial dimension. Our calculations reveal that accuracy of three significant digits for the energy can be attained by varying $\alpha_{1,i}$, while $\alpha_{2,i}$, $\alpha_{3,i}$, and $\alpha_{4,i}$ are set to infinity, which practically keeps the ground state (level 1) at its Hartree–Fock geometry. The optimal values of $\alpha_{1,i}$ for the ground level $|1\rangle$ are shown in Figure 2c for two, three, and four electrons, also showing that both $\alpha_{1,i}$ and the nonlocal length $\sigma_{1,i}$ are almost independent of the number of electrons—except for one electron at the ground state where there is no e–e interaction—and $\alpha_{1,i}$ is set to zero. The degree of entanglement for the four configurations of Figure 1a is quantified by the linear quantum entropy of Equation (18), as shown in Figure 2b. There are two distinct cases: In the first, the electrons are considered identical (blue line), to be compared with the result from the exact numerical solution of

the Schrödinger equation (green line). It can be seen that the linear entropy in this case remains almost constant, in close agreement with the exact numerical result. The second case, plotted with red dots in Figure 2b, depicts the linear entropy for the different electrons considered as distinguishable particles with the density matrix of Equation (17). It can be seen that the linear entropy for the distinguishable electrons increases due to the screening effect of the inner electrons, which causes larger fluctuations in the shape of the outer wave functions.

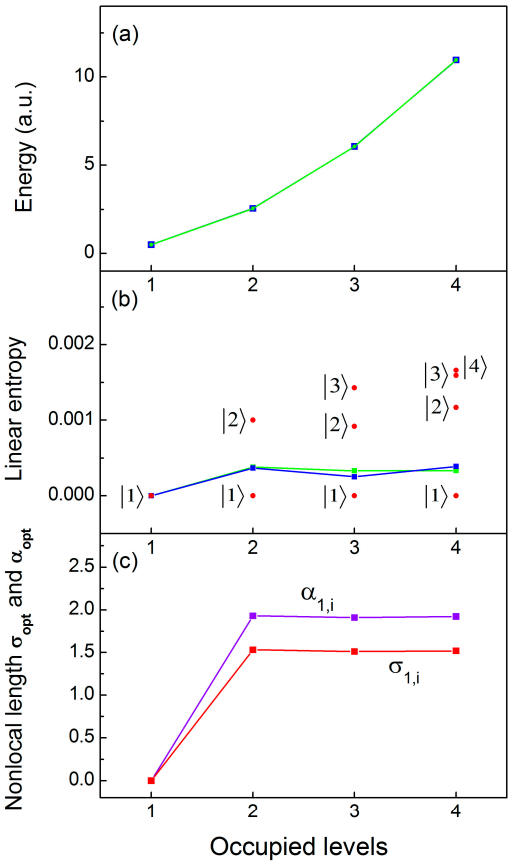


Figure 2. Energy (**a**), linear entropy (**b**), and nonlocality parameters $\sigma_{1,i}$ ($\alpha_{1,i}$) (**c**) for the ground state $|1\rangle$, for electron configurations with up to four parallel-spin electrons (Figure 1a). Blue lines: TDQMC results; green lines: numerically exact results; red dots in (**b**): linear entropy for distinguishable electrons.

For the filled-shell configuration of Figure 1b, each level contains two opposite-spin electrons, whose wave functions overlap in space almost completely, and therefore these electrons interact more like bosons [14]. It is therefore reasonable to calculate the only spatial entanglement that is due to same-shell states that is expected to significantly exceed the entanglement of the same-spin electrons at different levels. For the configuration of Figure 1b we have found that the major source of entanglement is the Coulomb repulsion between the two outermost electrons. For the ground state (level 1), where only two opposite-spin electrons are present in the vicinity of the core, the system energy exhibits a well-defined minimum as a function of the nonlocal length $\sigma^{\uparrow}_{1,1} = \sigma^{\downarrow}_{1,1}$, as seen in Figure 3a. When adding electrons at the higher levels, the corresponding wave functions acquire zeros, which is a source of larger fluctuations of the energy, as seen in Figure 3b–d, where the red curves represent the polynomial least squares fit for better visualization of the energy minimum.

Note that in the process of adding correlated electrons at the outer shells, the inner-shell electrons remain largely intact due to their stronger localization (confinement) to the core. Therefore, to a good approximation, the inner-shell wave functions need not be

recalculated, and these may remain at their self-consistent Hartree–Fock configurations. Figure 4a depicts the system energy (blue line), which almost perfectly matches the numerically exact energies (green line) obtained using the standard DMC method [10]. The linear entropy predicted by the TDQMC method decreases when adding new excited states to the electron configuration (blue line in Figure 4b), which contrasts with the case of bosonic quantum dots, where it increases for more electrons at the ground state [14]. This behavior is also confirmed by the numerically exact results (green line) for levels 1 and 2 (two and four identical electrons), and can be explained by the orthogonality of the wave functions in the fermionic calculation, which causes weaker interaction for the outer-shell electrons.

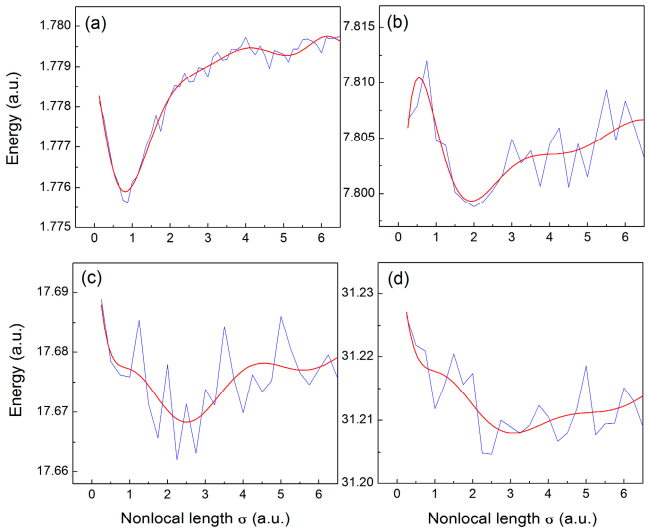


Figure 3. Energy of N-electron 1D quantum dots as a function of the nonlocal length $\sigma_{N,N}$, for N = 2 (**a**), N = 4 (**b**), N = 6 (**c**), and N = 8 (**d**).

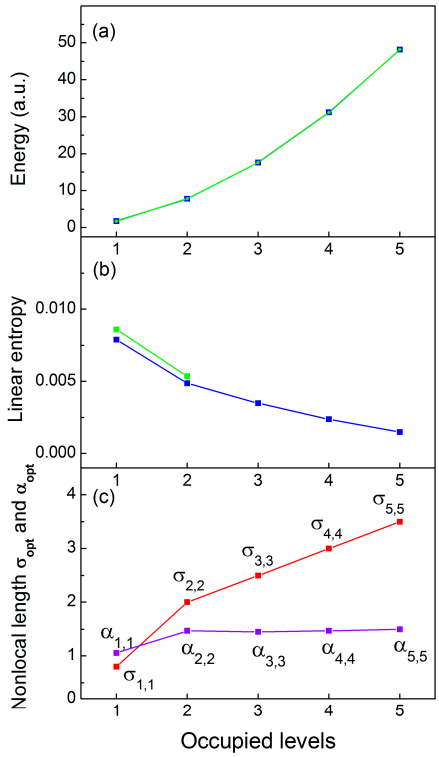


Figure 4. Energy (**a**), linear quantum entropy (**b**), and nonlocality parameters $\sigma_{1,i}$ ($\alpha_{1,i}$) (**c**) for the outermost level $|i\rangle$=$|1\rangle$, . . . ,$|5\rangle$-, for electron configurations with up to five filled shells (Figure 1b). Blue lines: TDQMC results; green lines: numerically exact results.

4. Conclusions

In conclusion, we have calculated the correlated ground state and excited states of 1D quantum dots with spin-polarized and spin-compensated electron configurations with up to 5 energy levels (10 electrons) within the time-dependent quantum Monte Carlo framework. Unlike the Hartree–Fock approximation, the TDQMC method accounts for the dynamic correlations between the constituents of the quantum system, and allows one to quantify the resultant spatial entanglement in a simple and efficient manner. By variationally minimizing the system energy with respect to the spatial quantum nonlocality, the optimal set of wave functions that describes each electron is found, which allows one further to calculate the reduced density matrices for the different electrons and, hence, to quantify the entanglement that they exhibit due to their mutual interactions. Using the linear quantum entropy as a measure for the entanglement, it was found that for a fully spin-polarized electron configuration the stochastic windowing applied to the ground state alone is sufficient to recover the entanglement of the excited states, in good agreement with the exact numerical result. For the spin-compensated electron system, the entanglement that is due to the interaction of the two outermost opposite-spin electrons is dominant, while the inner shells remain largely at their Hartree–Fock states. An essential advantage of this method is that it allows one to conceive quantum particles as identical as well as distinguishable objects. The theory presented here may find useful applications in treating quantum correlation effects in composite quantum systems such as molecules, clusters, and solid-state materials. Its accuracy could be further improved by using linear combinations of Slater determinants to better approach the Hilbert space of the quantum many-body problem.

Funding: This research is based on work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-7003, and by the Bulgarian Ministry of Education and Science as a part of National Roadmap for Research Infrastructure, grant number D01-401/18.12.2020 (ELI ERIC BG).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nielsen, M.A.; Chuang, I.L. *Quantum Computation and Quantum Information*; Cambridge University Press: Cambridge, UK, 2010.
2. Tichy, M.C.; Mintert, F.; Buchleitner, A. Essential entanglement for atomic and molecular physics. *J. Phys. B Mol. Opt. Phys.* **2011**, *44*, 192001. [CrossRef]
3. Amico, L.; Fazio, R.; Osterloh, A.; Vedral, V. Entanglement in many-body systems. *Rev. Mod. Phys.* **2008**, *80*, 517–527. [CrossRef]
4. Horodecki, R.; Horodecki, P.; Horodecki, M.; Horodecki, K. Quantum entanglement. *Rev. Mod. Phys.* **2009**, *81*, 865–942. [CrossRef]
5. Froese-Fischer, C. *The Hartree-Fock Method for Atoms: A Numerical Approach*; Wiley Interscience: New York, NY, USA, 1977.
6. Parr, R.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, NY, USA, 1989.
7. Alon, O.E.; Streltsov, A.I.; Cederbaum, L.S. Unified view on multiconfigurational time propagation for systems consisting of identical particles. *J. Chem. Phys.* **2007**, *127*, 154103. [CrossRef]
8. Szabo, A.; Ostlund, N. *Modern Quantum Chemistry*; Dover: New York, NY, USA, 1996.
9. Feiguin, A.E. The density matrix renormalization group. In *Strongly Correlated Systems*; Avella, A., Mancini, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 31–64.
10. Hammond, B.B.; Lester, W.; Reynolds, P. *Monte Carlo Methods in Ab Initio Quantum Chemistry*; World Scientific: Singapore, 1994.
11. Christov, I.P. Correlated non-perturbative electron dynamics with quantum trajectories. *Opt. Express* **2006**, *14*, 6906. [CrossRef] [PubMed]
12. Christov, I.P. Dynamic correlations with time-dependent quantum Monte Carlo. *J. Chem. Phys.* **2008**, *128*, 244106. [CrossRef] [PubMed]
13. Christov, I.P. Particle-wave dichotomy in quantum Monte Carlo: Unlocking the quantum correlations. *J. Opt. Soc. Am. B* **2017**, *34*, 1817. [CrossRef]
14. Christov, I.P. Spatial non-locality in confined quantum systems: A liaison with quantum correlations. *Few Body Syst.* **2020**, *61*, 45. [CrossRef]
15. Osenda, O.; Serra, P. Scaling of the von Neumonn entropy in a two-electron system near the ionization threshold. *Phys. Rev. A* **2007**, *75*, 042331–042339. [CrossRef]

16. Pont, F.M.; Osenda, O.; Toloza, J.H.; Serra, P. Entropy, fidelity, and double orthogonality for resonance states in two-electron quantum dots. *Phys. Rev. A* **2010**, *81*, 042518–042522. [CrossRef]
17. Nielsen, E.; Muller, R.P.; Carroll, M.S. Configuration interaction calculations of the controlled phase gate in double quantum dot qubits. *Phis. Rev. B* **2012**, *85*, 035319. [CrossRef]
18. Pham, D.N.; Bharadwaj, S.; Ram-Mohan, L.R. Tuning spatial entanglement in interacting two-electron quantum dots. *Phys. Rev. B* **2020**, *101*, 045306. [CrossRef]
19. Bransden, B.H.; Joachain, C.J. *Physics of Atoms and Molecules*; Longman: New York, NY, USA, 1982.
20. Holland, P.R. *The Quantum Theory of Motion*; Cambridge University Press: Cambridge, UK, 1993.
21. Christov, I.P. Time dependent spatial entanglement in atom-field interaction. *Phys. Scr.* **2019**, *94*, 045401. [CrossRef]
22. Christov, I.P. Electron-pair densities with time-dependent quantum Monte Carlo. *J. Atom. Mol. Phys.* **2013**, *2013*, 424570. [CrossRef]
23. Breuer, H.P.; Petruccione, F. *The Theory of Open Quantum Systems*; Oxford University Press: Oxford, UK, 2002.
24. Coe, J.P.; Sudbery, A.; D'Amico, I. Entanglement and density-functional theory: Testing approximations on Hooke's atom. *Phys. Rev. B* **2008**, *77*, 205122. [CrossRef]
25. Zanardi, P.; Zalka, C.; Faoro, L. Entangling power of quantum evolutions. *Phys. Rev. A* **2000**, *62*, 030301. [CrossRef]
26. Ghirardi, G.; Marinatto, L. General criterion for the entanglement of two indistinguishable particles. *Phys. Rev. A* **2004**, *70*, 012109. [CrossRef]
27. Plastino, A.R.; Manzano, D.; Dehesa, J.S. Separability criteria and entanglement measures for pure states of N identical fermions. *Europhys. Lett.* **2009**, *86*, 20005. [CrossRef]
28. Benavides-Riveros, C.L.; Toranzo, I.V.; Dehesa, J.S. Entanglement in N-harmonium: Bosons and fermions. *J. Phys. B At. Mol. Opt. Phys.* **2014**, *47*, 195503. [CrossRef]
29. Schliemann, J.; Ignacio Cirac, J.; Kus, M.; Lewenstein, M.; Loss, D. Quantum correlations in two-fermion systems. *Phys. Rev. A* **2001**, *64*, 022303. [CrossRef]
30. Buscemi, F.; Bordone, P.; Bertoni, A. Linear entropy as an entanglement measure in two-fermion systems. *Phys. Rev. A* **2007**, *75*, 032301. [CrossRef]
31. Grobe, R.; Eberly, J.H. Photoelectron spectra for two-electron system in a strong laser field. *Phys. Rev. Lett.* **1992**, *68*, 2905. [CrossRef] [PubMed]

Entanglement bounds on the performance of quantum computing architectures

Zachary Eldredge,[1,2] Leo Zhou,[3] Aniruddha Bapat,[1,2] James R. Garrison,[1,2] Abhinav Deshpande,[1,2] Frederic T. Chong,[4] and Alexey V. Gorshkov[1,2]

[1]*Joint Center for Quantum Information and Computer Science, NIST/University of Maryland, College Park, Maryland 20742, USA*
[2]*Joint Quantum Institute, NIST/University of Maryland, College Park, Maryland 20742, USA*
[3]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*
[4]*Department of Computer Science, University of Chicago, Chicago, Illinois 60637, USA*

(Received 24 October 2019; accepted 4 August 2020; published 26 August 2020)

There are many possible architectures of qubit connectivity that designers of future quantum computers will need to choose between. However, the process of evaluating a particular connectivity graph's performance as a quantum architecture can be difficult. In this paper, we show that a quantity known as the isoperimetric number establishes a lower bound on the time required to create highly entangled states. This metric we propose counts resources based on the use of two-qubit unitary operations, while allowing for arbitrarily fast measurements and classical feedback. We use this metric to evaluate the hierarchical architecture proposed by A. Bapat *et al.* [Phys. Rev. A **98**, 062328 (2018)] and find it to be a promising alternative to the conventional grid architecture. We also show that the lower bound that this metric places on the creation time of highly entangled states can be saturated with a constructive protocol, up to a factor logarithmic in the number of qubits.

DOI: 10.1103/PhysRevResearch.2.033316

I. INTRODUCTION

As the development of quantum computers progresses from the construction of qubits to the construction of intermediate-scale devices, quantum information scientists have increasingly begun to explore various architectures for scalable quantum computing [1–4]. Researchers have quantified the cost imposed by moving from one architecture to another [5,6] and optimized the placement of qubits on a fixed architecture [7–9]. Experimentalists have also begun to test different architectures in laboratory settings [10,11].

In this work, we are interested in developing tools to evaluate the relative performance of different architectures. Here, "architecture" refers to the connectivity graph that defines the allowable elementary operations between qubits. We propose a natural metric based on entanglement measures. When several physical models are represented by a graph $G = (V, E)$, with a set of vertices V corresponding to qubits, and a set of weighted edges E corresponding to two-qubit operations (where the weights denote the maximum rates of operations), a useful metric is given by what we dub the "rainbow time,"

$$\tau_{\mathrm{RB}}(G) = \max_{F \subset V, |F| \leqslant \frac{1}{2}|V|} \frac{|F|}{|\partial F|}, \tag{1}$$

where $|\partial F|$ denotes size of the boundary of F, i.e. the total weight of edges connecting F and $\bar{F} = V - F$.

We show that the rainbow time is a lower bound on the time required to create a highly entangled state on the graph (i.e., states of N qubits with $\mathcal{O}(N)$ bipartite entanglement). It is also the reciprocal of a well-studied graph quantity known as the isoperimetric number [12]. We note that this lower bound holds even when measurement and feedback are allowed to speed-up entanglement generation, such as in the case of Greenberger-Horne-Zeilinger states [13]. In contrast to Ref. [14], where architectures are evaluated assuming that only unitary operations are permitted, our results apply to the more general setting that allows nonunitary operations.

As a complementary result, we show that this lower bound is nearly tight—a procedure that distributes Bell pairs using maximum-flow algorithms nearly saturates this bound to produce $\mathcal{O}(N)$ entanglement across any bipartition, up to $\mathcal{O}(\log N)$ overhead. This suggests that beyond providing a bound, the rainbow time would be a useful witness to the speed at which entanglement can actually be generated.

II. PHYSICAL MODEL

In this paper, we evaluate the performance of quantum architectures with a connectivity graph given by G. Each vertex in the graph represents a single data qubit, and an edge exists between two vertices if two-qubit operations can be performed between them. We interpret the edge weight w_{ij} between vertices i and j as representing bandwidth, so that higher-weighted edges are capable of performing more two-qubit operations in a single unit of time.

We consider an example physical model where the edge weights represent the rate of distribution of entangled pairs

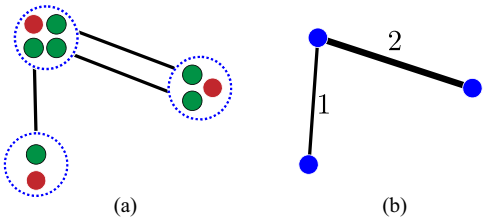


FIG. 1. Illustration of how a model with ancilla mediator qubits can be abstracted into one in which only data qubits and edge weights are tracked. In panel (a), each module (blue dashed circle) contains one data qubit (red) and several ancilla mediator qubits (green) that form Bell pairs with other modules. In panel (b), the module as a whole is represented by blue circles, while the ancilla mediator qubits are now represented by edge weights. Only the states of the data qubits are tracked.

as in Ref. [15]. Each vertex is a small module that contains a data qubit and some ancilla qubits. In each unit of time, Bell pairs are generated between the ancilla qubits on the edges of the graph, which can then be used to perform two-qubit gates on the data qubits [16,17]. The process of moving from this model to an abstracted connectivity graph is illustrated in Fig. 1. We assume that measurements, classical communication, and intra-module unitaries are arbitrarily fast, such that the bottleneck is given by quantum operations between modules. For example, this model can describe a trapped-ion system which uses photonic interconnects to generate entanglement between modules as in Refs. [18,19]. In this framework, vertex degrees and total graph edge weights represent required ancilla overheads, justifying their use as cost functions in Ref. [14].

While, for simplicity, we will focus in the main text on the above model, our results also apply to other physical models, up to constant-factor overheads. For example, since any two-qubit operation between data qubits can be performed by consuming two Bell pairs [20], the above model is equivalent to a model where edge weights are proportional to rates of two-qubit operations. In Appendix A, we show in more detail how to extend our results to this model, as well as to a model where edge weights represent coupling strengths in a Hamiltonian.

III. ENTANGLEMENT CAPACITY

Given a graph G, we wish to bound the total possible increase in a given entanglement measure after n rounds of entanglement distribution over its links. Suppose we fix a bipartition of the graph into two subgraphs supported on vertex subsets F and $\bar{F}$. We consider a general entanglement measure, S, which quantifies the bipartite entanglement between F and $\bar{F}$. We assume the following axioms: S is zero for product states $\rho_F \otimes \rho_{\bar{F}}$, additive between nonentangled regions, $S(\rho_{F\bar{F}} \otimes \tau_{F\bar{F}}) = S(\rho_{F\bar{F}}) + S(\tau_{F\bar{F}})$, and nonincreasing under local operations and classical communication. Entanglement measures that obey these axioms include the entanglement cost, the distillable entanglement, and the entanglement of formation [21,22]. All of these measures are identical to the von Neumann entropy for pure states.

By the result of Ref. [21], the entanglement after n rounds is bounded by n times the maximum single-round entanglement. We will therefore bound the entanglement generated in one round, going from ρ to ρ'. To produce ρ', we begin with ρ and then generate entanglement on the graph edges. This means that w_{ij} ancilla Bell pairs are generated for each edge (i, j) crossing the boundary ∂F. The total number of Bell pairs is therefore $|\partial F|$, the sum over all the weights,

$$|\partial F| = \sum_{i\in F, j\in\bar{F}} w_{ij}. \tag{2}$$

Ignoring ancillas purely local to F or $\bar{F}$, the resulting state is $\rho \otimes \rho_{\text{Bell}}^{\otimes|\partial F|}$. The final state ρ' is then generated by local operations, assisted by classical communication, on this state. We denote the state that results from an arbitrary round of local operations and classical communications on ρ as $\text{LOCC}(\rho)$. Therefore, our axioms for S allow us to write

$$\begin{aligned} S(\rho') &= S\big[\text{LOCC}\big(\rho \otimes \rho_{\text{Bell}}^{\otimes|\partial F|}\big)\big] \\ &\leqslant S\big(\rho \otimes \rho_{\text{Bell}}^{\otimes|\partial F|}\big) \\ &= S(\rho) + |\partial F| S(\rho_{\text{Bell}}), \\ &\Rightarrow S(\rho') - S(\rho) \leqslant |\partial F| S(\rho_{\text{Bell}}). \end{aligned} \tag{3}$$

Working in the units of $S(\rho_{\text{Bell}}) = 1$, we refer to this upper bound on the change in entanglement, $\Delta S \leqslant |\partial F|$, as the *entanglement capacity* of the $(F, \bar{F})$ bipartition in the graph G.

IV. RAINBOW STATES

We now define a highly entangled state whose creation serves as a benchmark for the performance of a quantum computing architecture.

Entanglement makes a useful benchmark for any quantum computer because it can be shown that computations that do not produce entanglement can be efficiently simulated classically [23–25]. Further motivation for producing highly entangled states can be found in quantum simulation, where a quantum simulator of general applicability ought to be capable of representing and simulating highly entangled states [26].

To select a particular entangled state for benchmarking, we consider "rainbow states." In 1D contexts, for even N, a rainbow state is one in which qubits i and $N - i$ are maximally entangled [27,28]. The state itself is maximally entangled across a bipartition between the first $N/2$ qubits and the rest.

We extend this construction to arbitrary graphs. Suppose we consider a set of qubits V and any subset $F \subset V$, with the requirement that $|F| \leqslant \frac{1}{2}|V|$. Denote by F_i the ith vertex of F using an arbitrary ordering, and similarly use $\bar{F}_i$ to index vertices in the complement $\bar{F}$. We can then define a "rainbow" state as one in which qubit F_i and qubit $\bar{F}_i$ form a Bell pair, and any additional qubits in $\bar{F}$ are left in the state $|0\rangle$. This state is illustrated for a particular choice of F and ordering in Fig. 2. Note that this construction is only well-defined if $|F| \leqslant \frac{1}{2}|V|$, as otherwise there will not be enough data qubits in $\bar{F}$ to form Bell pairs with all the data qubits in F. The arbitrary ordering allows multiple rainbow states to be defined from the same F.

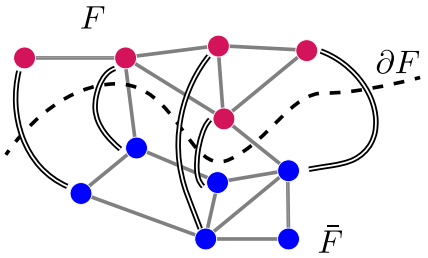


FIG. 2. An illustration of how a rainbow state is defined on an arbitrary subgraph F. Here, gray lines represent the connectivity graph of allowed two-qubit interactions, while doubled black lines represent maximally entangled qubit pairs. Qubits without a doubled line are assumed to be in state $|0\rangle$.

V. RAINBOW TIMES AND ISOPERIMETRIC NUMBER

Using the model for quantum architectures in which each edge weight of a graph G denotes the rate of entanglement generation across that edge, we can calculate the lower bound on the time required to create a rainbow state, according to the entanglement capacity. For any vertex subset F we define this time as

$$t(F) = \frac{|F|}{|\partial F|} = \frac{\text{number of qubits in } F}{\text{entanglement capacity of } (F, \bar{F})}. \tag{4}$$

As we have shown, the entanglement capacity corresponds to the total weight of edges across the boundary, which constrains the amount of entanglement that can be distributed to the subsystem F from its complement $\bar{F}$ in unit time.

Although there are many choices for a highly entangled physical state associated with the subset F that would be hard to create, here we argue why the above metric $t(F)$ suffices for most considerations. Although there are many different states with $\mathcal{O}(N)$ entanglement which could be used to evaluate graphs, the rainbow state is easy to conceptualize and create. Since any bipartite entangled state can be converted either to or from Bell pairs through entanglement concentration or dilution [29], the rainbow state offers insight into the time required to create a general bipartite entangled state. Furthermore, rainbow states arise as ground states of novel models in condensed-matter physics [30], and thus the ability to create them can be important for quantum simulation. The difficulty to create rainbow states is also recognized in Ref. [13]. While there is freedom in defining a physical rainbow state via the pairing of vertices in F with those in $\bar{F}$, the precise choice of pairing does not affect the minimum time required to create the state according to the entanglement capacity, $t(F)$. While different rainbow states that share a common subset F may differ in how quickly they can be created, $t(F)$ serves as the common lower bound on the creation time for all of them, and thus we will focus on that metric here.

We will now use $t(F)$ to evaluate the quantum architecture G, the larger graph that contains F as a vertex subset. To do this, we find the maximum $t(F)$ given G. Note that this is not the same as maximizing entanglement entropy, which would simply yield half the graph without any consideration of the graph structure. Instead we ask: Of all the maximally entangled states we can build by bipartitioning V into F and $\bar{F}$, which of them is slowest to build according to the entanglement capacity? We call the associated quantity $t(F)$ the *rainbow time* of the graph G and denote it $\tau_{\mathrm{RB}}(G)$, as defined in Eq. (1).

The rainbow time has a simple and attractive interpretation, can be directly connected to quantum computing tasks, and is applicable to various physical models of computation. In addition, it can be directly connected to a quantity known as the isoperimetric number $h(G)$ [12], sometimes also known as the Cheeger constant, which is well-studied in graph theory and computer science [31–33]. As we have defined it, the rainbow time is simply $\tau_{\mathrm{RB}}(G) = 1/h(G)$ [34]. Thus, aiming to minimize the rainbow time (so that large entangled states can be easily created) in a quantum architecture is equivalent to maximizing the isoperimetric number. An "isoperimetric set" is a vertex subset F that achieves $t(F) = \tau_{\mathrm{RB}}(G)$. Often, isoperimetric numbers appear in the context of expander graphs, which are constructed to possess large isoperimetric numbers [35] and are used to prove important results in complexity theory [36–38]. Intuitively, a small isoperimetric number (large τ_{RB}) means that a graph has bottlenecks, and a sizable subset can easily be disconnected by removing relatively few edges. This also implies that an architecture with large τ_{RB} is more prone to becoming disconnected due to the failure of a small number of edges.

Even though computation of the exact rainbow time is NP-hard for general graphs [12], it can be approximated to within an $\mathcal{O}(\sqrt{\log N})$ factor [39]. There are also efficiently computable bounds on the rainbow time, including ones using the eigenvalues of the graph Laplacian [12]. Furthermore, for many specific graphs, we can evaluate the rainbow time efficiently. In Appendix B, we have done this for the complete, star, and grid graphs, as well as the hierarchical products and hierarchies presented in Ref. [14]. In particular, we compare hierarchies to d-dimensional grids and show that, for some parameters, hierarchies have lower rainbow time and lower total edge weight than grids, making them promising architectures for quantum computing.

VI. CREATING RAINBOW STATES

So far we have shown that rainbow time τ_{RB} serves as a lower bound for generating maximum entanglement across any bipartition of the system. We now examine whether this bound can be saturated, in the sense that one can create a rainbow state across any bipartition in time $\tilde{\mathcal{O}}(\tau_{\mathrm{RB}})$. We will show that for a general graph, there is an explicit protocol that prepares a rainbow state in time no more than $\lceil \tau_{\mathrm{RB}} \ln |F| \rceil$ for any bipartition where F is the smaller subset, indicating that the bound τ_{RB} is tight up to a logarithmic factor.

We begin the proof by mapping the problem of creating rainbow state to the MaxFlow problem in computer science [40]. Here, we restrict our attention to quantum architectures on graph $G = (V, E)$, where the edge weights are integers that represent the number of Bell pairs that can be generated across the edge per unit time. Suppose we are given arbitrary vertex subsets F and K, where $|F| = |K| \leqslant |V|/2$, and $K \subset \bar{F}$. To create a Bell state between a given pair of nodes in a single time step, we can specify a path connecting them on the graph G, generate Bell pairs on each edge along that path, and then perform entanglement connection on each internal node to

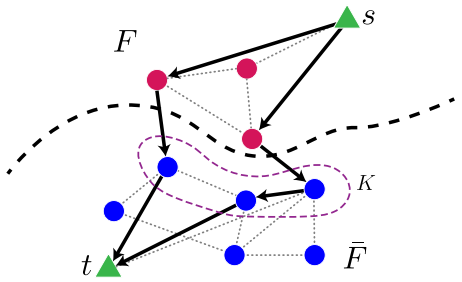


FIG. 3. An illustration of the fictitious nodes added to the isoperimetric set, F, and a set of equal size K (encircled by purple dashed line), to create a flow network. The new fictitious nodes, s and t, appear as green triangles connected to every node in F and K, respectively; the original nodes and edges are pink (in F) and blue (in $\bar{F}$) circles. The edges have weight one. The flow, shown by arrows, transfers $\lceil |F|/\tau_{\mathrm{RB}} \rceil = 2$ units of entanglement across the bipartition. Gray, dotted edges are not used by the flow.

convert the string of Bell pairs into one long-distance Bell pair. We can create many distant Bell pairs in this way during a single time step by specifying many paths. However, the set of paths must not use any edge more often than the weight of that edge allows for, since by definition the weight of an edge limits the number of Bell pairs the edge can generate in a unit time step. Thus, we can interpret the weight of each edge as its *capacity*, and the collection of paths as a *flow* of entanglement from F to K, as illustrated in Fig. 3. Suppose we now attach a fictitious source node s to each node in F, and a fictitious sink node t to every node in K. Then the problem of maximizing the number of Bell pairs simultaneously generated between F and K is the same as the problem of maximizing the flow from the source s to the sink t. The latter problem is known as MaxFlow, visualized in Fig. 3, and an explicit protocol to give the maximum possible amount of flow can be found efficiently via, e.g., the Ford-Fulkerson algorithm [41]. Note that if all the edge weights are integers, a flow of maximum value exists in which the flow carried by each edge is also an integer [42].

To demonstrate that a flow approach yields an efficient creation of a rainbow state, we invoke the MaxFlow-MinCut theorem, which says that the maximum flow has the same value as the minimum cut [40]. Here, a "cut" means a bipartition of the graph separating s and t, and its value is the total weight of all edges that cross the bipartition. By finding a lower bound on the value of all possible cuts in a graph, we show that a flow larger than or equal to this bound must exist.

Suppose that we now consider any cut of the graph into some arbitrary pair of subsets $\{s\} \cup S$ and $\{t\} \cup T$. The boundary of this cut will consist of edges from $s \to T$, $S \to t$, and $S \to T$. Its magnitude can be written as

$$|\mathrm{Cut}(S,T)| = |T \cap F| + |S \cap K| + |\partial S|, \tag{5}$$

since s and t are connected only to nodes in F and K, respectively, and the edges in $S \to T$ are just the boundary of S in the original graph. To evaluate $|\partial S| = |\partial T|$, we will assume that $|S| \leqslant \frac{1}{2}|V|$, meaning we can apply the isoperimetric condition $|S| \leqslant |\partial S|\tau_{\mathrm{RB}}$. (If this is not the case, then a near-identical argument can be made applying this condition to T.) To account for cases where $\tau_{\mathrm{RB}} < 1$, we will write this as $|\partial S| \geqslant m|S|$ where $m = \min(1, 1/\tau_{\mathrm{RB}})$. We then note that

$$\begin{aligned} |\partial S| &\geqslant m|S| \geqslant m(|S \cap F| + |S \cap K|) \\ &\geqslant m(|F| - |T \cap F| + |S \cap K|). \end{aligned} \tag{6}$$

By inserting this lower bound for $|\partial S|$ into Eq. (5), we obtain

$$|\mathrm{Cut}(S,T)| \geqslant (1-m)|T \cap F| + (1+m)|S \cap K| + m|F|. \tag{7}$$

Since we know $m \leqslant 1$, we obtain the final bound on the cut magnitude,

$$|\mathrm{Cut}(S,T)| \geqslant m|F|. \tag{8}$$

If $m = 1$ (i.e., $\tau_{\mathrm{RB}} \leqslant 1$), then it follows that the value of the smallest cut is greater than $|F|$, meaning that a flow exists of magnitude at least $|F|$, which creates the rainbow state in a single round. If $m < 1$ (i.e., $\tau_{\mathrm{RB}} > 1$), then we find that a flow exists of magnitude $|F|/\tau_{\mathrm{RB}}$ [43]. Once $|F|/\tau_{\mathrm{RB}}$ nodes are entangled, they can be disconnected from s and t, and the process repeated on a new set of nodes $F_1 \subset F$. Therefore, after n rounds of computation, the remaining set of nodes waiting for entanglement F_n is produced by removing $1/\tau_{\mathrm{RB}}$ of the nodes in set F_{n-1}, with $F_0 = F$, allowing us to compute the maximum size of F_n inductively:

$$\begin{aligned} |F_n| &\leqslant \left(1 - \frac{1}{\tau_{\mathrm{RB}}}\right)|F_{n-1}| \\ &\leqslant \left(1 - \frac{1}{\tau_{\mathrm{RB}}}\right)^n |F| < e^{-n/\tau_{\mathrm{RB}}}|F|. \end{aligned} \tag{9}$$

Once $|F_n| < 1$, the process is complete, as there are no fractional nodes. It follows that $\lceil \tau_{\mathrm{RB}} \ln |F| \rceil$ rounds suffice to complete the entangling process.

VII. OUTLOOK

In this work, we have presented a new metric for evaluating proposed architectures for quantum computers. While we have proven that any vertex subset F can have a rainbow state prepared in $\lceil \tau_{\mathrm{RB}} \ln |F| \rceil$ time, test simulations on many example small graphs suggest that flow-based algorithms can create rainbow states in $\lceil \tau_{\mathrm{RB}} \rceil$ time. It is thus possible that the logarithmic factor can be removed and that the rainbow time lower bound is fully tight and saturable. In addition, although our argument suggests that for any bipartition of the system, *there exists a* rainbow state that can be created in $\lceil \tau_{\mathrm{RB}} \ln |F| \rceil$ time, other rainbow states (where the connections between node pairs are permuted) may take longer. It would be interesting to upper bound the creation time of arbitrary rainbow states using tools from classical network theory such as routing time [44,45].

Finally, another open question is how the entanglement capacity, used here in terms of the rainbow time, can be applied to the analysis of quantum algorithms. While the rainbow time is not enough to provide an upper bound on the time-complexity of running a quantum algorithm on a given quantum architecture, it can provide a lower bound when the amount of entanglement required in the algorithm is known. References [46,47] explore the question of how entanglement grows during Shor's algorithm and in adiabatic quantum computing. These complement other results showing

that low-entanglement systems can be simulated efficiently on a classical computer [23,48]. Rainbow time can also be used to benchmark algorithms for compilation and gate decomposition of quantum circuits, by comparing their realized circuit depth to this theoretical minimum required time.

ACKNOWLEDGMENTS

We thank A. Childs, A. Harrow, L. Jiang, D. Leung, G. Smith, and X. Wu for discussions. Z.E., A.B., J.R.G., A.D., and A.V.G. acknowledge funding by ARO MURI, DoE ASCR Quantum Testbed Pathfinder program (Award No. DE-SC0019040), AFOSR, ARL CDQI, NSF PFCQC program, AFOSR MURI, DoE BES Materials and Chemical Sciences Research for Quantum Information Science program (Award No. DE-SC0019449), DoE ASCR Accelerated Research in Quantum Computing program (Award No. DE-SC0020312), and NSF PFC at JQI. A.B. is supported in part by the QuICS Lanczos Fellowship. Z.E. is supported in part by the ARCS Foundation. L.Z. is supported in part by the National Science Foundation and the Center for Ultracold Atoms. J.R.G. was supported in part by the NIST NRC Research Postdoctoral Associateship Award. F.C. is funded in part by EPiQC, an NSF Expedition in Computing, under grant CCF-1730449; in part by STAQ, under grant NSF Phy-1818914; and in part by DOE Grants No. DE-SC0020289 and No. DE-SC0020331. This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation Grant No. PHY-1607611.

APPENDIX A: ENTANGLEMENT CAPACITIES ON VARIOUS PHYSICAL MODELS

In this Appendix, we will derive the entanglement capacity for several different physical models that can correspond to a graph. Consider a graph, G, and select a subset of the vertices, F. We then want to show that the maximum amount of entanglement that can be created between F and $\bar{F}$ in unit time is proportional to the size of the boundary, $|\partial F|$. We will allow arbitrary constant factors, and discuss how this bound arises in two different physical situations. As in the main text, we consider entanglement measures S on two regions so long as S obeys the following rules:

(1) Additively distributive over the tensor product, so $S(\rho\otimes\sigma)=S(\rho)+S(\sigma)$ if ρ and σ are supported on both sides of the bipartition.

(2) Zero for states which are a product of states on each region, $S(\rho_F\otimes\rho_{\bar{F}})=0$.

(3) Nonincreasing after any operation which is local to each region, even if we permit classical communication.

In the main text, we showed how to apply these axioms to the analysis of a case in which computation was performed by the production and consumption of Bell pairs. Here we also look at a gate model of computation and a case in which the graph describes the limits on a time-dependent interaction Hamiltonian.

1. Unitaries

In this model, each graph edge of weight w_{ij} represents the capability to perform w_{ij} unitaries between qubits i and j in a time step. These unitaries are freely chosen by the experimenter. For two qubits, the ability to apply multiple unitaries is no different from the ability to apply an arbitrary unitary. However, we are considering cases where the qubits are part of a larger system, meaning we may wish to perform unitaries in sequence on different pairs to perform a more complicated computation.

We note that every two-qubit unitary can be performed using two Bell pairs as a shared resource and applying local operations. This can be easily seen in the following process:

(1) Alice and Bob start with a data qubit each and two Bell pairs shared between them. They wish to implement an arbitrary two-qubit unitary using only local operations and classical control.

(2) Alice uses one Bell pair and classical communication to teleport her qubit to Bob.

(3) Bob uses his local operations to perform the desired two-qubit gate.

(4) Bob teleports Alice's qubit back to her.

Therefore, the state ρ' can be obtained from the state ρ by using local operations and classical communication (LOCC) and consuming up to $2|\partial F|$ Bell pairs in the process. Since LOCC cannot increase S, it follows that

$$S(\rho')\leqslant S\left(\rho\otimes\rho_{\text{Bell}}^{\otimes 2|\partial F|}\right) \tag{A1}$$

$$\Rightarrow \Delta S\leqslant 2|\partial F|S(\rho_{\text{Bell}}). \tag{A2}$$

This suggests that the ability to perform arbitrary unitaries is up to twice as powerful as the ability to distribute arbitrary Bell pairs, which makes sense, as an arbitrary two-qubit gate cannot necessarily be performed with one Bell pair (for instance, SWAP requires two) [20]. Two Bell pairs, however, suffice to implement any arbitrary two-qubit unitary. In any case, this still yields an entanglement capacity $\Delta S=\mathcal{O}(|\partial F|)$ bound as desired.

2. Hamiltonians

We will now consider a case in which the graph describes a Hamiltonian, possibly time-dependent. The graph will restrict the strength of these Hamiltonians. If we assume that $G=(V,E)$, then the Hamiltonian can be written as a sum over the two-qubit operations:

$$H(t)=\sum_{(i,j)\in E}h_{ij}(t). \tag{A3}$$

We then impose the condition

$$\forall t:\|h_{ij}(t)\|\leqslant w_{ij}, \tag{A4}$$

where w_{ij} is the i-j edge weight. We can then apply the "small incremental entangling" (SIE) theorem [49]. In particular, we apply the special case used in Ref. [50] to bound the total amount of entanglement generated by this Hamiltonian. If H is a sum of pairwise Hamiltonians h_{ij} acting on qubits, then the time-rate of entanglement generation on a set F of sites is

$$\left|\frac{dS_F}{dt}\right|\leqslant 36\log(2)\sum_{i\in F,j\in\bar{F}}\|h_{ij}\|. \tag{A5}$$

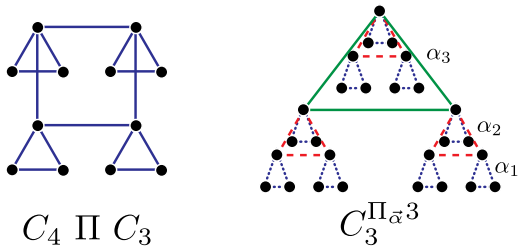


FIG. 4. Examples of a hierarchical product (left) and a weighted hierarchy (right).

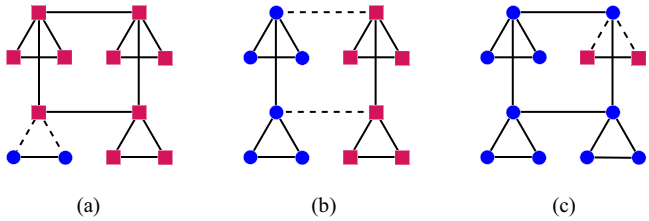


FIG. 5. Three classes of subgraph used in our proof. Circles represent vertices in F, squares are vertices in $\bar{F}$, and dashed lines are edges in ∂F. (a) A situation in which part of one copy of H_2 is in F. (b) A situation in which the division between F and $\bar{F}$ lies entirely in H_1. (c) A situation in which all but one of the copies of H_2 are entirely contained in F.

Here, S_F is the von Neumann entropy of the reduced density matrix on the region F. This can be derived from Eq. (3) of Ref. [50], and specifying two-body terms and qubit sites, but the result could be extended to qudits or general k-body interactions. The sum over Hamiltonian norms, in the graph context, corresponds to a sum over graph edges. Since every Hamiltonian strength is limited by the corresponding edge weight, $\sum \|h_{ij}\| \leqslant \sum w_{ij} = |\partial F|$. Therefore, we can specifically say that for this case, $\Delta S_F = \mathcal{O}(|\partial F|)$. Many other entanglement measures, such as entanglement of formation or entanglement cost, can be related to the von Neumann entropy [22]. In particular, many entanglement measures on mixed states can be defined as a weighted sum over pure state components; since none of the pure states can increase dramatically in entanglement under this process, the entanglement measure on the mixed state is similarly limited.

APPENDIX B: APPLICATION TO HIERARCHICAL PRODUCT AND HIERARCHIES

In this Appendix, we calculate the rainbow times for the hierarchical products and hierarchies of Ref. [14]. A hierarchical product is a graph product denoted $G \,\Pi\, H$ in which $|G|$ copies of H are connected at their root (first) vertices by the graph G. By iterating this process, we can create a hierarchy, in which higher-level graphs connect lower-level identical subhierarchies. We also extend this concept to that of a weighted hierarchy, in which the edges on level i have weight α_i. We write a k-level hierarchy with a vector of weights $\vec{\alpha}$ as $G^{\Pi_{\vec{\alpha}} k}$, where G is the base graph. Finally, if $\alpha_i = \alpha^{i-1}$, so that edge weight scales geometrically with the level of the hierarchy, then we simply write $G^{\Pi_\alpha k}$. Some examples are shown in fig. 4.

To calculate the rainbow time for a hierarchical product, we make use of the result from Ref. [12] that there must exist an isoperimetric set [a vertex set F such that $\tau(F) = \tau_{\mathrm{RB}}(F)$] that is connected and whose complement $\bar{F}$ is connected. Therefore, we will look at all possible subgraphs of $H_1 \,\Pi\, H_2$ where both F and $\bar{F}$ are connected. From these, we will search for the one with the largest $\tau(F)$. Since some isoperimetric set is guaranteed to exist in this set of subgraphs, this maximization over $\tau(F)$ in this set will also give us $\tau_{\mathrm{RB}}(H_1 \,\Pi\, H_2)$. We will begin by specifying three cases, illustrated in fig. 5. These cases cover all possible subsets with the right connectedness properties and therefore allow us to find the maximizing set for the graph and $\tau_{\mathrm{RB}}(H_1 \,\Pi\, H_2)$.

One such set would cover part of one copy of H_2. However, note that if the root vertex of H_2 were included in F, then we would have to include all the descendants of H_2, since otherwise $\bar{F}$ would not be connected. Therefore, this class will only include subsets of H_2 which do not include the root vertex. In this case, we must maximize over all possible subsets of H_2 to find the maximum $\tau(F)$. This may seem like it would yield $\tau_{\mathrm{RB}}(H_2)$; however, in this instance we can pick subsets of H_2 which make up a majority of H_2, which is not allowed for τ_{RB}. We define the *unrestricted rainbow time* as

$$u_{\mathrm{RB}}(G) = \sup_{F \subset G \setminus G_1} \tau(F). \tag{B1}$$

Here, $G \setminus G_1$ refers to G with its first vertex removed. Therefore, any set from this class will offer a candidate rainbow time of at most $\tau(F) = u_{\mathrm{RB}}(H_2)$.

The second class of candidate sets would cross one or more copies of H_2. Since F must be connected, the path between these copies must be included in F, which means the root vertices of each H_2 that connect to each other via H_1 must also be in F. Then, as shown above, the entire copy of H_2 must be included. As a result, this case is equivalent to choosing copies of H_2 and either entirely including them in F or entirely excluding them. This problem reduces to dividing up H_1, and then calculating as if each vertex had an effective volume of $|H_1|$. Therefore, we can find the maximum $\tau(F)$ of these sets by simply finding $\tau_{\mathrm{RB}}(H_1)$ and scaling it by $|H_2|$.

The final class of sets F which meets the connectedness criteria would be an F which includes all of H_1 and then all but one copy of H_2 completely, with perhaps some of the remaining H_2 also included. However, this F would necessarily be larger than half of the total graph $H_1 \,\Pi\, H_2$, and therefore we can discard it as a candidate set for determining the rainbow time. We combine the first two options and conclude that

$$\tau_{\mathrm{RB}}(H_1 \,\Pi\, H_2) = \max\left[u_{\mathrm{RB}}(H_2), |H_2| \tau_{\mathrm{RB}}(H_1)\right]. \tag{B2}$$

We now seek to apply this to hierarchies $G^{\Pi_{\vec{\alpha}} k}$. Just as before, if a vertex is included in F, then we must also include in F all its descendants in the hierarchy; otherwise, the complement $\bar{F}$ will not be connected. Therefore, all bipartitions can be reduced to choosing a particular level of the hierarchy to cut—on that level, either a vertex will be included or not included, and this must apply to all of its descendants as well. Every bipartition can then be mapped to a bipartition of G, but one where every vertex is scaled by $|G|^{i-1}$ due to the size of

each subhierarchy [note that the large number of vertices not in F do not contribute to $\tau(F)$]. In addition, $\tau(F)$ must also be modified by the edge weight, which we define to be α_i on level i.

There is one important difference between the top (kth) level and all others, which arises from the constraint that $|F| \leqslant \frac{1}{2}|G^{\Pi_{\vec{\alpha}}k}|$. A cut on the top level must not include more than half of the highest-level copy of G, while all lower levels can use any cut at all as long as it does not include the root vertex. Whatever level we cut, the cut depends only on the base graph G, with each node standing for $|G|^{i-1}$ total nodes below it. Therefore, we can write the overall τ_{RB} as a maximization over these options:

$$\tau_{\mathrm{RB}}(G^{\Pi_{\vec{\alpha}}k}) = \max\left(\frac{|G|^{k-1}}{\alpha_k}\tau_{\mathrm{RB}}(G),\ \sup_{i<k}\frac{|G|^{i-1}}{\alpha_i}u_{\mathrm{RB}}(G)\right). \tag{B3}$$

For specificity, we will evaluate the case where $G = K_n$, the complete graph, and $\alpha_i = \alpha^{i-1}$, which was proposed in Ref. [14] as an architecture. Here, the maximization over lower levels [the second term in Eq. (B3)] can be reduced to either to the first level or the $k-1$ level, since we simply have to pick the largest element in a geometric sequence defined by n/α. We can write the resulting maximization as a choice between three options,

$$\tau_{\mathrm{RB}}\left(K_n^{\Pi_\alpha k}\right) = \max\left[1,\ \left(\frac{n}{\alpha}\right)^{k-1}\frac{2}{n},\ \left(\frac{n}{\alpha}\right)^{k-2}\right]. \tag{B4}$$

Whereas one might have expected two options to arise (cut at the top or at the bottom), we actually have three. For $\alpha > n$, the edges grow in capacity too quickly for the increased volume to make a higher-level cut worthwhile, so the optimal cut is at the bottom, yielding a constant scaling with n. Two other options appear at $n > \alpha$, where cutting higher up the hierarchy allows for greater volume of qubits in F without too much penalty caused by changing edge weights. The reason there are two strategies is that it may be possible to cut a larger portion of a lower hierarchy and exploit the split between τ_{RB} and u_{RB}. [For K_n, in particular, the cut that includes all but the root vertex satisfies $u_{\mathrm{RB}}(K_n)$.]

TABLE I. Important statistics for graphs. Here, only the asymptotic scaling with N is written. In addition to the rainbow time τ_{RB} for each graph, we also include the total weight of all edges w, and the maximum graph degree Δ. Rainbow times for graphs other than hierarchies can be found in terms of isoperimetric number in Refs. [12,32].

Graph Name	τ_{RB}	w	Δ
K_N	N^{-1}	N^2	N
S_N	1	N	N
d-dimensional Grid	$N^{1/d}$	N	$2d$
$K_n^{\Pi_\alpha k}$	$N^{\max(0,1-\log_n\alpha)}$	$N^{\max(1,\log_n\alpha)}$	$\log_n N$

To place these results in context, we compare the rainbow time of $K_n^{\Pi_\alpha k}$ to the total rainbow time of other graphs. To do this, we write the rainbow time in terms of the total number of qubits in a graph, N, and concern ourself with the overall scaling. For the purpose of comparison, we consider hierarchies where the number of levels scale logarithmically as $k = \log_n N$, while α, n are constant parameters independent of N. In this language, $\tau_{\mathrm{RB}}(K_n^{\Pi_\alpha k}) = \Theta(N^{\max(0,1-\log_n\alpha)})$. We compare this to the rainbow time of some other graphs in Table I. References [12,32] give the isoperimetric number for K_N, S_N (the star graph of N nodes) and grids (which are Cartesian products of paths). Satisfying sets for these graphs are: for K_N and S_N, an arbitrary half of the nodes; for grids, a hypercube placed in one corner that takes up half the total volume.

One goal would be to identify a set of parameters where a hierarchy outperforms a d-dimensional grid architecture. We are most concerned with comparing to the d-dimensional grid because the other candidates we present, K_N and S_N, both have very large degree, making them impractical for scalable architectures, although both have been used for small quantum devices [10]. We find that the rainbow time of the hierarchy with base graph K_n and scaling constant α will be better (smaller) than that of the grid if $\alpha > n^{(d-1)/d}$. If it also holds that $n > \alpha$, then the hierarchy will accomplish this with a total edge weight scaling identically as the grid. It is possible to achieve a smaller prefactor in this scaling under a suitable choice of n, α; for example, when $d = 2$, the choice of $n = 3, 4$ and $\alpha = n^{1-1/d}$ gives lower total edge weight for the hierarchy than the grid. We conclude that a hierarchy $K_n^{\Pi_\alpha k}$ with $\alpha \in [n^{1-1/d}, n)$ has both lower rainbow time and lower total edge weight than a d-dimensional grid of qubits.

[1] C. Monroe and J. Kim, Science **339**, 1164 (2013).

[2] M. Ahsan and J. Kim, in *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE'15)* (IEEE Conference Publications, Washington, D.C., 2015), pp. 1108–1113.

[3] A. Pirker, J. Wallnöfer, and W. Dür, New J. Phys. **20**, 053054 (2018).

[4] B. Villalonga, S. Boixo, B. Nelson, C. Henze, E. Rieffel, R. Biswas, and S. Mandrà, npj Quantum Inf. **5**, 86 (2019).

[5] D. Cheung, D. Maslov, and S. Severini, in *Proceedings of the Workshop on Quantum Information* (2007).

[6] A. Holmes, S. Johri, G. G. Guerreschi, J. S. Clarke, and A. Y. Matsuura, Quantum Sci. Technol. **5**, 025009 (2020).

[7] D. Rosenbaum and M. Perkowski, in *Proceedings of the 40th IEEE International Symposium on Multiple-Valued Logic* (IEEE, Washington, D.C., 2010), pp. 270–275.

[8] D. J. Rosenbaum, in *Proceedings of the 8th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC'13)*, Leibniz International Proceedings in

Informatics (LIPIcs), Vol. 22, edited by S. Severini and F. Brandao (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2013), pp. 294–307.
[9] M. Pedram and A. Shafaei, IEEE Circuits Syst. Mag. **16**, 62 (2016).
[10] N. M. Linke, D. Maslov, M. Roetteler, S. Debnath, C. Figgatt, K. A. Landsman, K. Wright, and C. Monroe, Proc. Natl. Acad. Sci. USA **114**, 3305 (2017).
[11] D. Maslov, New J. Phys. **19**, 023035 (2017).
[12] B. Mohar, J. Combin. Theory, Ser. B **47**, 274 (1989).
[13] C. Meignant, D. Markham, and F. Grosshans, Phys. Rev. A **100**, 052333 (2019).
[14] A. Bapat, Z. Eldredge, J. R. Garrison, A. Deshpande, F. T. Chong, and A. V. Gorshkov, Phys. Rev. A **98**, 062328 (2018).
[15] K. S. Chou, J. Z. Blumoff, C. S. Wang, P. C. Reinhold, C. J. Axline, Y. Y. Gao, L. Frunzio, M. H. Devoret, L. Jiang, and R. J. Schoelkopf, Nature **561**, 368 (2018).
[16] D. Gottesman and I. L. Chuang, Nature **402**, 390 (1999).
[17] L. Jiang, J. M. Taylor, A. S. Sørensen, and M. D. Lukin, Phys. Rev. A **76**, 062323 (2007).
[18] K. R. Brown, J. Kim, and C. Monroe, npj Quantum Inf. **2**, 16034 (2016).
[19] R. Nigmatullin, C. J. Ballance, N. de Beaudrap, and S. C. Benjamin, New J. Phys. **18**, 103028 (2016).
[20] J. Eisert, K. Jacobs, P. Papadopoulos, and M. B. Plenio, Phys. Rev. A **62**, 052317 (2000).
[21] C. H. Bennett, A. W. Harrow, D. W. Leung, and J. A. Smolin, IEEE Trans. Inf. Theory **49**, 1895 (2003).
[22] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, Rev. Mod. Phys. **81**, 865 (2009).
[23] G. Vidal, Phys. Rev. Lett. **91**, 147902 (2003).
[24] F. Verstraete, J. J. García-Ripoll, and J. I. Cirac, Phys. Rev. Lett. **93**, 207204 (2004).
[25] Although universal quantum computation is possible in the limit of vanishing entanglement by implementing any quantum circuit $\mathcal{C}$ in a way that's controlled by a qubit in the state $\sqrt{1-\epsilon}|0\rangle + \sqrt{\epsilon}|1\rangle$ [51], such computation still requires the ability to implement the circuit $\mathcal{C}$. This means that any entanglement-based bound on the time-complexity of implementing $\mathcal{C}$ would still apply to the ϵ-entangled version.
[26] J. I. Cirac and P. Zoller, Nat. Phys. **8**, 264 (2012).
[27] G. Ramírez, J. Rodríguez-Laguna, and G. Sierra, J. Stat. Mech. (2015) P06002.
[28] R. N. Alexander, A. Ahmadain, Z. Zhang, and I. Klich, Phys. Rev. B **100**, 214430 (2019).
[29] C. H. Bennett, H. J. Bernstein, S. Popescu, and B. Schumacher, Phys. Rev. A **53**, 2046 (1996).
[30] Z. Zhang, A. Ahmadain, and I. Klich, Proc. Natl. Acad. Sci. USA **114**, 5142 (2017).
[31] B. Mohar, Linear Alg. Appl. **103**, 119 (1988).
[32] F. R. K. Chung and P. Tetali, Comb. Probab. Comput. **7**, 141 (1998).
[33] F. Chung, Ann. Comb. **9**, 1 (2005).
[34] Note that $\tau_{\mathrm{RB}}(G)$ can take on any nonnegative real value. In reality, the creation of a quantum state will always take an integer number of steps greater than or equal to one in our model. Therefore, $\lceil \tau_{\mathrm{RB}} \rceil$ can be used as a measure of the "number of rounds required" in cases where this is important.
[35] O. Goldreich, in *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, Lecture Notes in Computer Science (Springer, Berlin, 2011), pp. 451–464.
[36] M. Ajtai, J. Komlós, and E. Szemerédi, in *Proceedings of the 15th Annual ACM Symposium on Theory of Computing (STOC'83)* (ACM, New York, NY, 1983), pp. 1–9.
[37] O. Reingold, J. ACM **55**, 17 (2008).
[38] I. Dinur, J. ACM **54**, 12 (2007).
[39] S. Arora, S. Rao, and U. Vazirani, J. ACM **56** (2009).
[40] P. Elias, A. Feinstein, and C. Shannon, IRE Trans. Inf. Theory **2**, 117 (1956).
[41] L. R. Ford and D. R. Fulkerson, Can. J. Math. **8**, 399 (1956).
[42] D. R. Fulkerson and Rand Corporation, *Notes on Linear Programming. Part XVL, A Network-Flow Feasibility Theorem and Combinatorial Applications*, ASTIA Document No. AD 156011 (Rand Corporation, Santa Monica, CA, 1958).
[43] Since it is guaranteed to be integer-valued for graphs with integer-valued edge weights, the flow must in fact be of magnitude $\lceil |F|/\tau_{\mathrm{RB}} \rceil$, but this makes no difference to the argument.
[44] E. Schoute, L. Mancinska, T. Islam, I. Kerenidis, and S. Wehner, arXiv:1610.05238.
[45] A. M. Childs, E. Schoute, and C. M. Unsal, in *Proceedings of the 14th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC'19)*, Leibniz International Proceedings in Informatics (LIPIcs), Vol. 135, edited by W. van Dam and L. Mancinska (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019), pp. 3:1–3:24.
[46] R. Orús and J. I. Latorre, Phys. Rev. A **69**, 052308 (2004).
[47] V. M. Kendon and W. J. Munro, Quantum Inf. Comput. **6**, 630 (2006).
[48] N. Schuch, M. M. Wolf, K. G. H. Vollbrecht, and J. I. Cirac, New J. Phys. **10**, 033032 (2008).
[49] K. Van Acoleyen, M. Mariën, and F. Verstraete, Phys. Rev. Lett. **111**, 170501 (2013).
[50] Z.-X. Gong, M. Foss-Feig, F. G. S. L. Brandão, and A. V. Gorshkov, Phys. Rev. Lett. **119**, 050501 (2017).
[51] M. Van den Nest, Phys. Rev. Lett. **110**, 060504 (2013).

Entanglement at a scale and renormalization monotones

Published for SISSA by Springer

Received: October 28, 2018
Revised: December 31, 2018
Accepted: January 22, 2019
Published: January 29, 2019

Entanglement at a scale and renormalization monotones

Nima Lashkari

Center for Theoretical Physics, Massachusetts Institute of Technology,
77 Massachusetts Avenue, Cambridge, MA 02139, U.S.A.

E-mail: lashkari@mit.edu

Abstract: We study the information content of the reduced density matrix of a region in quantum field theory that cannot be recovered from its subregion density matrices. We reconstruct the density matrix from its subregions using two approaches: scaling maps and recovery maps. The vacuum of a scale-invariant field theory is the fixed point of both transformations. We define the entanglement of scaling and the entanglement of recovery as measures of entanglement that are intrinsic to the continuum limit. Both measures increase monotonically under the renormalization group flow. This provides a unifying information-theoretic structure underlying the different approaches to the renormalization monotones in various dimensions. Our analysis applies to non-relativistic quantum field theories as well the relativistic ones, however, in relativistic case, the entanglement of scaling can diverge.

Keywords: Field Theories in Higher Dimensions, Nonperturbative Effects, Renormalization Group

ArXiv ePrint: 1704.05077

https://doi.org/10.1007/JHEP01(2019)219

JHEP01(2019)219

Contents

1 Introduction 1
- 1.1 Measuring asymmetry 2
- 1.2 Measuring non-Markovianity 2

2 Entanglement of scaling 4

3 Markov states in QFT 6

4 Entanglement at a scale 10

5 Renormalization monotones 11

6 Conclusions 12

A The entanglement of scaling is monotonic 13

JHEP01(2019)219

1 Introduction

In recent years, the techniques and intuitions from quantum information-theory have proven to be immensely helpful in the study of many-body quantum systems. The entanglement structure of the low energy states of local Hamiltonians is a key concept in simulating lattice systems in condensed matter, the study of order parameters in phase-transitions, and constructing renormalization monotones in relativistic quantum field theories.

The renormalization group (RG) flow is the process in which one integrates out the ultraviolet (UV) high energy degrees of freedom, and compensates for them by adjusting the coupling constants such that the low energy physics is unchanged. Since the information about the UV modes are washed out, one might expect that the RG flow is irreversible. RG monotones are functions that reflect this irreversability as they change monotonically under the flow.

The study of RG monotones in relativistic quantum field theory (QFT) was started by the seminal work of Zamolodchikov [1], where he showed that the two point function of stress tensor in $2d$ QFT is a monotonic function of scale. In four dimensions, it was conjectured by Cardy in [2], and later proved in [3], that the a-anomaly term is an RG monotone. In two and three dimensions, the strong subadditivity (SSA) of entropy was used to show that there are universal terms in the entanglement entropy of vacuum in QFT reduced to a ball-shaped region that are RG monotones [4]. At the moment, the approaches to construct RG monotones seem to depend on the dimensionality of the spacetime, and a framework that works for all dimensions is missing.

In field theory, scaling is a unitary operation that allows us to compare the reduced density matrices on subsystems of different size. In this paper, we use scaling and the recovery maps of quantum information theory to quantify the amount of long-range quantum correlations at a scale. As a crucial step, we show that the Markov property of the vacuum of a conformal field theory implies that the vacuum state reduced to a null cone can be recovered perfectly from its subregions using both maps. We define the entanglement of scaling and the entanglement of recovery as two measures whose first derivative quantifies the *long-range entanglement*.[1] Both of these functions increase monotonically under the RG flow. In some relativistic theories the entanglement of scaling can be infinite; however, we expect that the entanglement of recovery to remain finite. Our monotonic functions are generalizations of the $2d$ and $3d$ entanglement monotones to higher dimensions. They provide a unifying information-theoretic approach to RG monotones in various dimensions. Furthermore, it points to a connection between recovery maps in quantum information theory and the RG transformation of states that goes beyond the construction of monotones.[2] We start by reviewing some notions and tools in quantum information theory.

1.1 Measuring asymmetry

Consider a many-body finite quantum system split into n non-overlapping regions A_1 to A_n, with isomorphic Hilbert spaces on A_i. The relabeling of the subsystem index i is a unitary operation in the global Hilbert space: $\otimes_{i=1}^{n}\mathcal{H}_i$. A simple example of such a unitary is the translation defined by $i \to i+1 \mod n$:

$$U = \sum_{a_1\cdots a_n} |a_2\cdots a_n a_1\rangle\langle a_1\cdots a_n|,$$

where $\{a_i\}$ is the basis that spans $\mathcal{H}_i$. The density matrix ρ_i on A_i is mapped to A_{i+1} with the local unitary

$$\begin{aligned}\rho_{i+1} &= \mathcal{E}(\rho_i) = U_i^\dagger \rho_i U_i \\ U_i &= \sum_{a_i,a_{i+1}} |a_{i+1}\rangle\langle a_i|. \end{aligned} \tag{1.1}$$

If the transformation sends a subsystem A to $\tilde{A}$, and the state is asymmetric under this transformation, some information about ρ_A will be lost. The relative entropy $S(\rho_{\tilde{A}}\|\mathcal{E}(\rho_A))$ is a measure of the amount of information in ρ_A that is lost. It is non-negative, and vanishes if and only if ρ_A is symmetric under the transformation.

1.2 Measuring non-Markovianity

Imagine that we are probing the global state with detectors that are localized in A_1A_2. The von Neumann entropy $S(\rho_{12})$ is a measure of the amount of quantum information ρ_{12} is missing about a pure global state. If we made a larger detector that allows us access to the

[1] Intuitively, we think of the entanglement of scaling to be a generalization the measure introduced in [22] to general non-relativistic field theories.

[2] While this manuscript was in preparation, the papers [5, 6] appeared, which have overlaps with some results presented here.

JHEP01(2019)219

region $A_1A_2A_3$, then the new detector teaches us $S(A_3|A_1A_2)$ more qubits of information. The quantity $S(A|A') \equiv S(AA') - S(A)$ is the conditional entropy. Another way to gain more information is by moving our detectors to adjacent sites A_2A_3. This gives us access to both ρ_{12} and ρ_{23}; however, we are still missing the *long-range* correlations between A_1 and A_3. We would like to quantify the amount of quantum information ("entanglement") about in ρ_{123} that is neither in ρ_{12} nor in ρ_{23}. Naively, one can say that by moving the detector we have learned $S(A_3|A_2)$ but there are still

$$I(A_1 : A_3|A_2) \equiv S(A_3|A_1A_2) - S(A_3|A_2) \tag{1.2}$$

more qubits in ρ_{123} that we are missing. This quantity is the conditional mutual information (CMI), and is non-negative by the SSA inequality [7].

A careful study of the operational question of how well can one guess ρ_{123} from the knowledge of ρ_{12} and ρ_{23} (the marginals) suggests that this naive estimate (CMI) is, indeed, a good measure of the amount of long-range entanglement. This can be seen from the two arguments below:

1. Statistical physicist's prescription for the best guess is to consider the set of all consistent global states $\mathcal{C}$; that is all ϕ_{123} with $\phi_{12} = \rho_{12}$ and $\phi_{23} = \rho_{23}$. The best guess is a state ϕ_{123} in this set, which has the largest entropy [8]. It follows from the consistency condition that the entropy of the best guess is the CMI:

$$\sup_{\phi_{123}\in\mathcal{C}} S(\phi_{123}) = I(A_1 : A_3|A_2). \tag{1.3}$$

2. Quantum information theorist's approach is to look at *recovery* maps. If a state has zero CMI, it can be reconstructed perfectly from its marginals. Such states are called quantum Markov states, and satisfy the following property:

$$\log \phi_{123} = \log \phi_{12} + \log \phi_{23} - \log \phi_2. \tag{1.4}$$

   The Markov state has no genuine long-range quantum correlations. All the correlations between A_1 and A_3 is classical and conditioned on A_2 [9]. Furthermore, when the CMI is small one can use universal recovery maps to reconstruct the global state with high fidelity [10, 11]. The CMI provides an upper bound on the fidelity distance of the recovered state. In fact, if we do not require the recovery map to be a quantum channel one can write down the explicit map

$$\rho_{\text{recov}} = e^{\log \rho_{12} + \log \rho_{23} - \log \rho_2}/Z, \tag{1.5}$$

   that is hardly distinguishable from the global state:

$$S(\rho_{123}|\rho_{\text{recov}}) \leq I(A_1 : A_3|A_2). \tag{1.6}$$

   Here Z is the normalization of the state. The inequality above is satisfied trivially because $Z \leq 1$ [12].

JHEP01(2019)219

In our n-partite A_1 to A_n example, if the state ρ_{123} is Markovian one can recover it perfectly from ρ_{12} and ρ_{23}, move the detector to the an adjacent site, and try to recover ρ_{1234} fromρ_{123} and ρ_{34}. This can be iterated to reconstruct $\rho_{1\cdots m}$ for any $m < n$. If the state is recovered perfectly at each step, the global state is called a Quantum Markov chain [13, 14]. A quantum Markov chain found from adjacent local density matrices of size r has the form

$$\log \rho_{1\cdots,m+r} = \log \rho_{m\cdots,m+r} + \sum_{k=1}^{m} (\log \rho_{k\cdots,k+r-1} - \log \rho_{k+1\cdots,k+r-1}). \tag{1.7}$$

In our terminology, these Markov states have no entanglement at any scale larger than r.

Intuitively, a quantum Markov chain is scale-invariant, in the sense that all the information in a density matrix of size R can be recovered perfectly from subsystems of size $r < R$. This suggests that quantum Markov states should appear naturally as the fixed points of the renormalization group flow.

2 Entanglement of scaling

The states of a quantum field theory are wavefucntionals of fields: $\Psi(\phi(x))$. The transformations $f : x^\mu \mapsto x^\mu + \xi^\mu$ (diffeomorphisms) are the generalization of the relabeling operation in finite systems to the continuum limit. Analogously, diffeomorphisms act on the global state as unitary operators: $|\tilde{\psi}\rangle = e^{i\int d\Sigma^\mu \xi^\nu T_{\mu\nu}}|\psi\rangle$, where Σ is the spacelike surface where the state lives, and $T_{\mu\nu}$ is the stress tensor. If we split the degrees of freedom into a subregion A and the complement, then the unitary operator that maps the reduced state on A to the reduced state to $\tilde{A}$ is:

$$U = \int [D\phi]_g |(f^{-1})^*\phi\rangle\langle\phi| \tag{2.1}$$

where $(f^{-1})^*$ is the pull-back of functions from A to $\tilde{A}$ [15].

A familiar example of such diffeomorphisms is the generalization of translations in finite systems to the continuum limit. In quantum field theory, the translations are described by the unitaries $U = e^{ia^\mu P_\mu}$ which map ρ_A to $\tilde{\rho}_{\tilde{A}}$:

$$\langle \phi_a(x_\in A)|\rho_{A,g}|\phi_b(x \in A)\rangle = \langle (f^{-1})^*\phi_a|\rho_{\tilde{A},\tilde{g}}|(f^{-1})^*\phi_b\rangle,$$

where $\tilde{g} = (f^{-1})^* g$ is the transformed metric. If the translation is a symmetry of the background metric, and the state then the density matrix changes only by a unitary rotation.

In the remainder of this work, we will be interested in how local Dilatations acts on null cones. In polar coordinates, this maps $f : (t, r) \mapsto (e^{\lambda(\Omega)}t, e^{\lambda(\Omega)}r)$, and leaves the perpendicular directions Ω untouched; see figure 1. Take a ball on the time slice $t = R$ centered at $r = 0$. The boundary of this ball is on the null cone defined by $r - t = 0$. The dilatation f with constant λ rescales the size of the ball from R to $e^\lambda R$, and moves it from $t = R$ to $t = e^\lambda R$. The metric transforms by an overall conformal factor: $\tilde{g} = e^{2\lambda} g$. If the state is scale-invariant, for instance the vacuum of a scale-invariant theory, one can ignore

JHEP01(2019)219

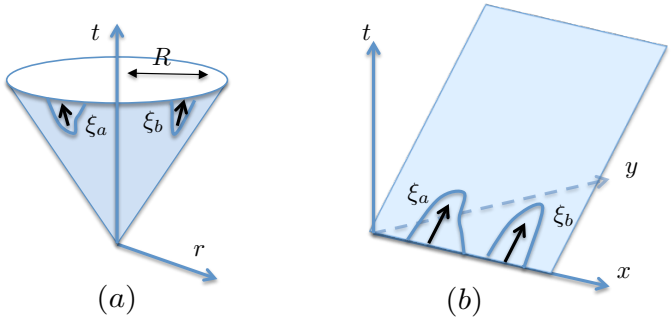


Figure 1. (a) Dilatataions that deform the boundary of ball at $t = R$, and act locally at particular angular variables Ω_a and Ω_b (b) Translations in the null direction that act locally in x coordinates.

the change of the metric, and the state remains unchanged up to a unitary. To simplify the notation, we denote the unitarily scaled density matrix from R to R' by

$$\tilde{\rho}_{R'} \equiv \mathcal{E}(\rho_R) = U^\dagger \rho_R U, \tag{2.2}$$

where R' has been suppressed in the notation, and will be clear from the context.

We are interested in a quantum field theory that is a deformation of a scale-invariant theory by a relevant operator of scaling dimension $\Delta < d$

$$S_{\rm QFT} = S_{\rm scale-inv} + \lambda_0 \int d^d x \mathcal{O}(x), \tag{2.3}$$

where $\lambda_0 = \mu^{\Delta - d} g_0$ is the dimensionful coupling at the UV length scale μ. Diffeomorphism invariance allows us to compare ρ_R, the reduced states on a ball of size R, to a smaller ball ρ_r rescaled back to R. In the UV ($r/\mu \ll 1$), the state ρ_r can be approximated well by the scale-invariant vacuum state which transforms trivially under rescaling $\mathcal{E}$. In essence, the entanglement of scaling compares the reduced density matrix of a QFT to that of its ultra-violet fixed point. with corrections proportional to the coupling λ_0. The modular operator of ρ_r can be computed in the conformal perturbation theory. It remains local in spacetime, to the first order in λ_0. The relative entropy $S(\rho_R \| \mathcal{E}(\rho_r))$ is a measure of the amount of distinguishability lost under the dilatation. We define the *entanglement of scaling* to be

$$\mathcal{S}_{sc}(\rho_R) = \lim_{r \to 0} S(\rho_R \| \mathcal{E}(\rho_r)). \tag{2.4}$$

The entanglement of scaling is, by definition, non-negative. Similar to the entanglement entropy, the entanglement of scaling is invariant under any unitary operations: $\mathcal{S}_{sc}(\rho) = \mathcal{S}_{sc}(U^\dagger \rho U)$.

In essence, the relative entropy above compares the reduced density matrix of quantum field theory with that of its fixed point which was proposed as a C-function in relativistic quantum field theories in [22]. As the authors of [22] have discussed, this measure can be divergent in relativistic QFT for deformations that are not relevant enough.

JHEP01(2019)219

3 Markov states in QFT

Take a quantum field theory density matrix ρ_R. If it is a quantum Markov state,[3] it can be perfectly recovered from its smaller marginals ρ_r, for any $r < R$. This suggests that there is no new physics at any length scale in between the r and R. In other words, it is scale-invariant in that range. One might expect that the CFT vacuum reduced to ball-shaped regions are quantum Markov states. In this section, we show that this intuition is indeed correct.

Start with a ball-shaped region A in a CFT vacuum state, and make two geometric deformations f_a and f_b. The state will be Markovian if the CMI $I(\delta A_a, \delta_b A|A)$ vanishes for any finite size deformation. This quantity was computed in a perturbation theory in small deformations by [16]. They find the CMI to be

$$I(\delta A_a; \delta A_b|A) = \delta A_a^{\bar{i}} \delta A_b^{(j)} \frac{2\pi^2 C_T}{(d+1)R^2} \frac{\eta_{\bar{i}\bar{j}}}{|\Omega_a - \Omega_b|^{2(d-1)}}, \tag{3.1}$$

where $\eta_{\bar{i}\bar{j}}$ and $\delta A_a^{(i)}$ and $\delta A_b^{(j)}$ are, respectively, the metric and the area elements in the t, r directions, and C_T is the coefficient in the two-point function of the stress tensor. For a generic deformation, this CMI is non-zero. However, if we take the deformed ball to be on a null cone, that is $\xi = \xi^u(\Omega)\partial_u$, the CMI is proportional to η_{uu} which is zero in flat space. This leaves the possibility that for null deformations the vacuum state is Markovian. This was recently proved to be case in [5]. Here, we explore the Markov property from an intuitive tensor network point of view using the method of the Euclidean path-integrals. In fact, it is pedagogical to start with a simpler example:

Ex. 1: QFT vacuum on half-space. As the first example, we show that the QFT vacuum in flat space reduced to a half-space is a quantum Markov state with respect to null deformations; see figure 1. Consider the vacuum of a $d > 2$ dimensional QFT in flat space $ds^2 = dudv + dx^2 + dz_i dz^i$, with $u = y + t$ and $v = y - t$ the null directions. We reduce the state to the region A, the $y > 0$ half-space. The modular operator of this region, $K_A \equiv -\log \rho_A$, is local [17]. On the null surface $v = 0$, it has the form

$$\begin{aligned} K_A &\equiv -\log \rho_A = \int dx K_x \\ K_x &= \int d^{d-3}z \int_0^\infty du\, u T_{uu}(x). \end{aligned} \tag{3.2}$$

In Euclidean QFT, the density matrix ρ_A is represented by a path-integral on $\mathbb{R}^d$, with boundary conditions above and below A in the Euclidean time; i.e. $(\tau_E = 0^\pm, y > 0)$ [18]. One can split the x direction into n slabs $A_i = (x_i, x_{i+1})$, and insert the resolutions of identity in between slabs; see figure 2:

$$\begin{aligned} \rho &= \int \prod_{i=1}^{N} [D\phi_i]\, \rho_i(\phi_i, \phi_{i+1}), \\ \rho_i(\phi_i, \phi_{i+1}) &= \langle \phi_i | \rho_i | \phi_{i+1} \rangle. \end{aligned} \tag{3.3}$$

[3] In the remainder of this paper, we use the words Markov chain and Markov states synonymously.

JHEP01(2019)219

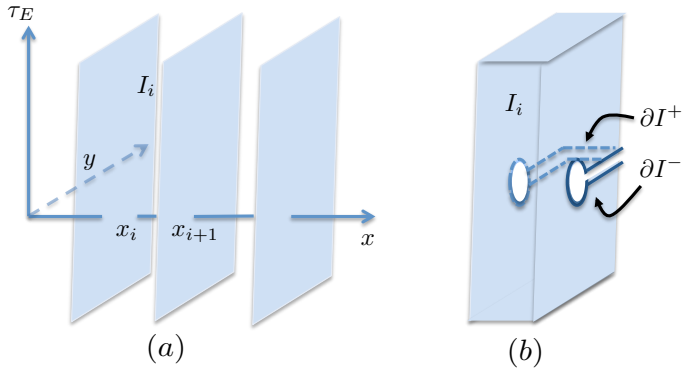


Figure 2. (a) Partitioning the Euclidean path-integral into slabs in the x directions (b) The path-integral over each slab has five boundaries. Two boundaries at x_i and x_{i+1}, two at ∂I^+ and ∂I^- where the state lives, and one infinitesimal cylinder cut around the origin at $y = \tau_E = 0$.

Here, $\rho_i(\phi_i, \phi_{i+1})$ is an operator (transfer matrix) that acts only on the subsystem A_i. Intuitively, one can think of the expression in (3.8) as a matrix product operator in the x direction; see figure 4.

We apply a diffeomorphism that is non-zero only at A_a and A_b, and deforms A to $\tilde{A} = A + \delta_a A + \delta_b A$. The density matrix of $\tilde{A}$ is given by $\rho_{\tilde{A},\eta} = U^\dagger \rho_{A,g} U$, where $g_{\mu\nu} = \partial_\mu \xi_\nu + \partial_\nu \xi_\mu + \partial_\mu \xi_\alpha \partial_\nu \xi^\alpha$, and η is the flat metric [15]. We take f to be a translation in a null direction localized on two slabs I_a and I_b:

$$f_a : u \mapsto u + \lambda f(x_a), \tag{3.4}$$

with $f(x_a)$ a function that has a peak at the center of A_a, and goes to zero on the boundaries of I_a at x_a and x_{a+1}.[4] The flat metric changes by $g_{xv} = \partial_x \xi_v = \lambda \partial_x f(x_a)$, which is nonzero only inside the slab I_a and vanishes on the boundaries ∂I_a. Partitioning the path-integral of $\tilde{\rho}_{\tilde{A}}$ according to (3.3) and comparing with ρ_A, only the transfer matrices ρ_a and ρ_b have changed. Let us focus on the matrix elements of one of these operators, $\tilde{\rho}_a$:

$$\langle \phi^1(\partial I_a^-) | \tilde{\rho}_a(\phi_a, \phi_{a+1}) | \phi^2(\partial I_a^+) \rangle = \int_{\phi(x_a)=\phi_a, \phi(\partial I_a^-)=\phi^1}^{\phi(x_{a+1})=\phi_{a+1}, \phi(\partial I_a^+)=\phi^2} [D\phi] e^{-S[\phi,g]}, \tag{3.5}$$

where $\partial I_a^\pm$ are the boundaries at $x \in A_a$ and $\tau_E = 0^\pm$; see figure 2. The path-integral above is on I_a that has five boundaries in the Euclidean $\mathbb{R}^{d+1}$. Two boundaries at $x = x_a$, $x = x_{a+1}$, two boundaries at ∂I_a^+ and ∂I_a^-, and a fifth boundary at $y^2 + \tau_E^2 = \epsilon$ which is a small cylinder cut around $y = \tau_E = 0$.

The only difference between the path-integrals for $\tilde{\rho}_a$ and ρ_a is in the metric that goes into the action. We Taylor expand the action around the flat space

$$S[\phi, g] = \exp\left(\int_{I_a} \partial^\mu \xi^\nu \frac{\delta}{\delta g^{\mu\nu}} \right) S[\phi, \eta] = \exp\left(- \int_{I_a} \xi^\nu \partial_\mu \frac{\delta}{\delta g^{\mu\nu}} + \int_{\partial I_a} d\Sigma^\mu \xi^\nu \frac{\delta}{\delta g^{\mu\nu}} \right) S[\phi, \eta], \tag{3.6}$$

[4] One might worry about the fact that the function f is not infinitely differentiable. We will be ignorant of such subtleties here.

JHEP01(2019)219

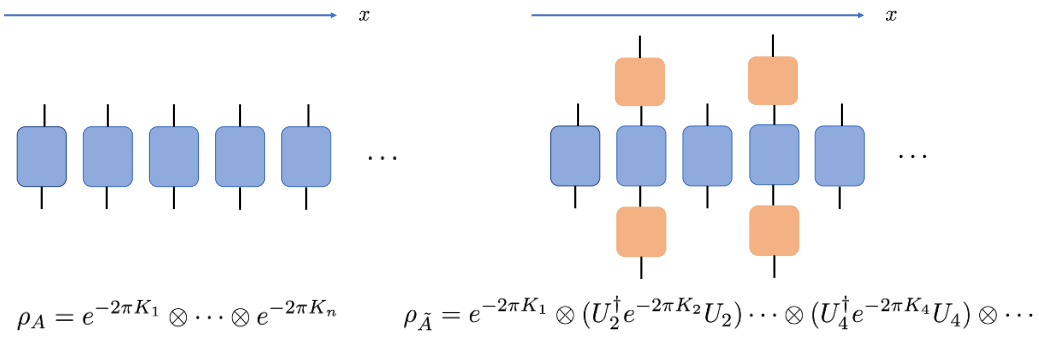


Figure 3. (a) The density matrix of the half-space on a null sheet factorizes in free field theory (b) a shape deformation on the null sheet at point $x = a$ corresponds to acting with unitaries U_a.

where we have used the integration by parts, and $d\Sigma^\mu$ is the normal to the boundary ∂I_a. The term with the integral over I_a vanishes, due to the fact that $\partial_\nu \frac{\delta}{\delta g^{\nu\mu}} S[\phi, g] = \partial_\nu T^{\mu\nu}$, which is identically zero.

The change in the metric under the diffemorphism by f_a is in the g^{ux} component, and since ξ^μ has only u components, only the two boundaries at constant x contribute to (3.6). However, we chose ξ to vanish on these boundaries; therefore $S[\phi, g]$ on I_a can be replaced with its flat space value $S[\phi, \eta]$. Hence, the transfer matrices in the partitioned path-integral in (3.3) do not change:

$$\tilde{\rho}_a(\phi_a, \phi_{a+1}) = \rho_a(\phi_a, \phi_{a+1}). \tag{3.7}$$

Hence, there is a unitary that rotates the overall density matrix ρ_A to $\tilde{\rho}_{\tilde{A}}$:

$$\tilde{\rho}_{\tilde{A}} = (\mathbb{I} \otimes U_a^\dagger \otimes U_b^\dagger) \rho_A (\mathbb{I} \otimes U_a \otimes U_b)\,. \tag{3.8}$$

This unitary operator is $U_a(x) = e^{i\alpha Q_a}$ where $Q_a = \int du\, T_{uu}(a)$ is the average null energy operator.

In the null quantization of free field theory, the vacuum state is the zero eigenvector of the null momentum P_u. Furthermore, we know that this state is a tensor product of the vacuua of the Q_x:

$$|\Omega\rangle = \otimes_x |\Omega_x\rangle, \qquad Q_x |\Omega_x\rangle = 0\,. \tag{3.9}$$

This means that the reduced density matrix of half-space is also a tensor product

$$\rho = \otimes_x \rho_x = \otimes_x e^{-2\pi K_x} \tag{3.10}$$

where ρ_x is the vacuum density matrix on the half-space found from the ground state $|\Omega_x\rangle$. There is no entanglement between ρ_x and $\rho_{x'}$ and the matrix product operator is of the form in figure 3. It is clear that applying the unitaries U_a and U_b only changes the matrices ρ_a and ρ_b and cannot create entanglement. Therefore, it is trivially true in free theory that

$$K_{\tilde{A}} = K + (U_a^\dagger K_a U_a - K_a) + (U_b^\dagger K_b U_b - K_b).$$

The two-dimensional Poincare group gives us the commutation relation

$$[K_x, Q_a] = -iQ_a \delta(x - a)\,. \tag{3.11}$$

JHEP01(2019)219

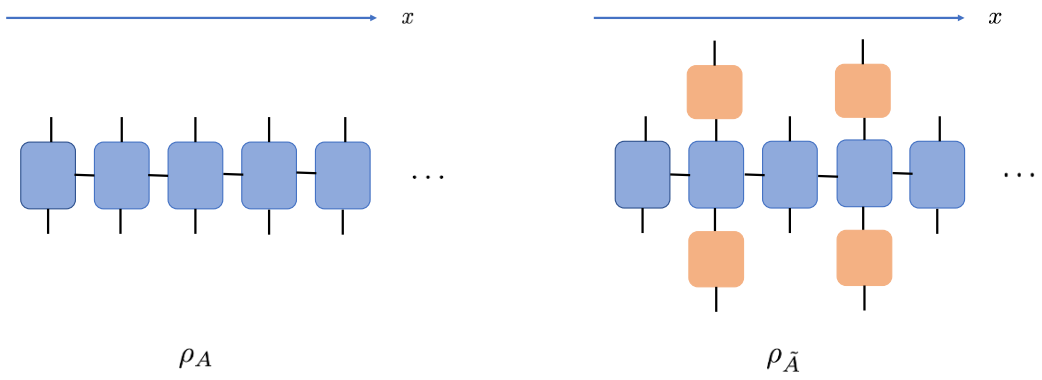


Figure 4. (a) The density matrix of the half-space on a null sheet factorizes in interacting theories is entangled in the x direction (b) a shape deformation on the null sheet at point $x = a$ corresponds to acting with unitaries U_a.

which results in a resummation of the Baker-Campbell-Hausdorff expansion:

$$U_x^\dagger e^{-2\pi K_x} U_x = e^{-2\pi(K_x - \alpha Q_x)} \,. \tag{3.12}$$

As a result, the modular Hamiltonian of the deformed region is

$$K_{\tilde{A}} = K_A - \alpha(Q_a - Q_b) \,. \tag{3.13}$$

This is the Markov property of vacuum in free field theory as was originally argued for in [19].

In a general interacting theory the vacuum state is the zero eigenvector of Q_x smoothed in the x direction. However, we expect Q_x with no smoothing to have no normalizable zero eigenvector.[5] This is reflected in the fact that the vacuum state is entangled across cuts of constant x. The matrix product operator representation of the vacuum density matrix is schematically drawn in figure 4. The density matrix is still

$$\rho_A = e^{-2\pi K_1} e^{-2\pi K_2} \cdots e^{-2\pi K_n} \tag{3.14}$$

which is not a product state. It has been argued in [5] that the commutator

$$[K_x, Q_a] = -iQ_a \delta(x - a) \,. \tag{3.15}$$

remains unmodified in interacting theories. One can commute the operators $e^{i\alpha Q_x}$ with $e^{-2\pi K_{x'}}$ and finds the same expression for the modular Hamiltonian as in the free theory:

$$K_{\tilde{A}} = K_A - \alpha(Q_a - Q_b), \tag{3.16}$$

which is the Markov property of the vacuum density matrix on a null sheet.

[5] We thank Juan Maldacena for pointing this out to us.

JHEP01(2019)219

Ex. 2: CFT vacuum on a null cone. There is a conformal transformation that maps the causal development of a half-space A to the causal development of a ball B [17]. If K_A and K_B are, respectively, the modular operators of subsystems A and B, there exists a unitary such that $K_B = U^\dagger K_A U$. Under this conformal transformation, the deformed half-space $A+\delta_a A$ is mapped to a deformed ball $B+\delta_a B$; see figure 1. Deformations on the null surface in A are sent to deformations of B on the null-cone. The equation (3.13) with $\tilde{A}$ continues to hold for the vacuum of a CFT in arbitrary dimensions with $\tilde{A}$ a deformation of the ball on the null cone that is its causal development. As a result, the vacuum of a d-dimensional CFT is a quantum Markov state with respect to deformations on a null cone.

In 2d CFTs, any state that is a descendant of vacuum with arbitrary time-dependence is related to vacuum by a conformal transformation, and remains a quantum Markov state. It is straightforward to check that SSA is saturated in these states from the expressions in [20].[6]

Near Markov states. Before applying the SSA inequality to the states of a quantum field theory, we would like to have an analogue of CMI that is insensitive to the ultraviolet details. We replace the entanglement entropies in CMI with the entanglement of scaling:

$$\begin{aligned} I_{sc}(A_1 : A_3|A_2) &\equiv \mathcal{S}_{sc}(\rho_{12}) + \mathcal{S}_{sc}(\rho_{23}) - \mathcal{S}_{sc}(\rho_2) - \mathcal{S}_{sc}(\rho_{123}) \\ &= I_{\rho_R}(A_1 : A_3|A_2) - \lim_{r\to 0} I_{\rho_r}(A_1 : A_3|A_2) \\ &= I(A_1 : A_3|A_2) \geq 0, \end{aligned} \tag{3.17}$$

where we have used the fact that the UV CFT state is Markovian. Note that in relativistic quantum field theory there is no guarantee that this quantity remains finite term by term.

4 Entanglement at a scale

In this section, for simplicity we restrict to vacuum state of QFTs in flat space.[7] The goal is to find an information-theoretic measure that quantifies the entanglement at a scale that is insensitive to the UV and has an operational interpretation. A measure of entanglement at scale R is a function that ρ_R and its derivatives $\partial_R^m \rho_R$. Here, we compare three candidate measures that appear natural from an information-theory point of view:

1. The obvious candidate is the relative entropy $S(\rho_{R+\delta R}|\mathcal{E}(\rho_R))$. This quantity vanishes at the first order in δR, due to the smoothness of relative entropy. At the second order, it becomes the quantum Fisher information which is a metric in the space of density matrices:

$$S(\rho_{R+\delta R}|\rho_R) = (\delta R)^2 \langle \delta^R \rho, \delta^R \rho \rangle_R + O((\delta R)^3).$$

   It is finite, non-negative at any R, and vanishes in CFTs. It is a metric, and hence satisfies the triangle inequality. Quantum Fisher information has an interpretation in terms of distinguishability, as it is the variation of a relative entropy.

[6] We thank Matthew Roberts for pointing this out to us.

[7] The generalization of the measures introduced here to arbitrary states requires minor, but straightforward modifications.

JHEP01(2019)219

2. The second candidate is the derivative $\partial_R \mathcal{S}_{sc}(\rho_R)$. It is finite, and non-negative at any R (see the supplementary material for a proof):

$$\partial_R \mathcal{S}_{sc}(R) \geq 0. \tag{4.1}$$

This quantity is expected to be insensitive to the UV details, and has the benefit that its integral, $\mathcal{S}_{sc}$, resembles a smoothed-out version of $S_{\rm UV} - S_{\rm IR}$. However, in relativistic field theory it diverges for deformations that are not relevant enough.

3. The third candidate, the information-theorist's favorite, is based on recovery maps and SSA. The task is to quantify how well one can recover the state $\rho_{R+\delta R}$ from the knowledge of all balls of size R within the causal development of $\rho_{R+\delta R}$. That is to say, we want to build a ball of size $R+\delta R$ from the iteration of a recovery map which acts on balls of size R. One way to do this was introduced in [4]. Take two balls with boundaries on a null cone. As we bring the balls close in the angular directions on the cone,the distance between $\delta_a A$ and $\delta_b A$ tends to R. the CMI measures the entanglement at scale R. To obtain the larger $\rho_{R+\delta R}$ we have to apply the recovery map many times following [4], and add up the CMI contributions at each step. The total sum of the CMI we obtain as we repeat this recipe is the quantity that we define to be the derivative of the *entanglement of recovery*

$$\partial_R \mathcal{S}_{\rm rec}(\rho_R) \equiv \left((d-3)\partial_R + R\partial_R^2\right) \mathcal{S}_{sc}(R) \geq 0. \tag{4.2}$$

It is a measure of the entanglement in the vacuum of QFTs at the scale R, that has an operational interpretation in terms of recovery. It vanishes in a CFT vacuum. Integrating this quantity from the UV to the scale R we obtain

$$\mathcal{S}_{\rm rec}(R) = (d-2-R\partial_R)\mathcal{S}_{sc}(\rho_R). \tag{4.3}$$

5 Renormalization monotones

We are encouraged by [21] to look for an RG monotone in arbitrary dimensions that has the following properties

1. It is a finite dimensionless quantity, and regularization independent.
2. It decreases monotonically along the flow.
3. If the flow ends in an IR fixed point, the value of the function can only depend on quantities that are intrinsic to the UV and IR fixed points.

We expect both the entanglement of scaling and the entanglement of recovery to satisfy the first property in non-relativistic examples. In relativistic theories, the conditions under which they remain finite is unclear to us and deserves further study. Both measures satisfy the second criterion:

$$\begin{aligned} \partial_R S_{sc}(R) &\geq 0 \\ \partial_R S_{\rm rec}(R) &\geq 0. \end{aligned} \tag{5.1}$$

JHEP01(2019)219

In all the known examples in $2d$ and $3d$ they also satisfy the third criterion. It is unclear to us, whether this continues to be the case in all dimensions.

In $2d$ and $3d$ they do indeed reduce to all the known monotones. The entanglement of scaling, $\mathcal{S}_{sc}(R)$, is a smoothed version of the RG monotone defined in [22], which is the relative entropy of vacuua in two different CFTs. While intuitive, the smoothness of $\mathcal{S}_{sc}(R)$ deserves further investigation. We believe that studying the entanglement of scaling in more detail can shed light on the UV divergences in the quantity in [22] for the particular range of the deformation scaling dimensions $\Delta > (d+2)/2$.

The entanglement of recovery, $\mathcal{S}_{\text{rec}}(R)$, is a smoothed version of the entanglement monotones in 2d and 3d introduced in [4] generalized to arbitrary dimension. As this work was in its final stages, we learned about the work in [6] that generalizes the previous entanglement proof to the a-theorem in four dimensions. It is of great interest to relate the entanglement of recovery to other known quantities of CFTs in $d > 4$.

6 Conclusions

In this work, we studied a connection between recovery maps in quantum information theory, and the renormalization group flow in quantum field theories. Applying information-theoretic tools, and taking advantage of the diffeomorphism invariance of QFT, we constructed candidate functions for the entanglement at a scale. Two new entanglement measures intrinsic to the continuum limit, the entanglement of scaling and the entanglement of recovery were defined. They are built such that their first derivatives in scale quantifies the amount of entanglement at scale. However, the more natural quantity from the point of view of the recovery maps is the entanglement of recovery. Both quantities are monotonic under a change of scale. A better understanding of the RG monotones in higher dimensions can be achieved by studying these quantities and relating them to the properties of the IR scale-invariant fixed point.

It is tempting to rewrite the entanglement of scaling in the language of the algebraic QFT as

$$\lim_{\lambda\to 0}\langle\Omega|\Delta_{\Omega,U_\lambda^\dagger\Omega U_\lambda}|\Omega\rangle, \tag{6.1}$$

and avoid referring to the density matrix. Here, $|\Omega\rangle$ is the state of a QFT, and $\Delta_{\Omega,\Omega'}$ is the relative modular operator of the two states with respect to a region, and U_λ generates dilatation by factor λ. We postpone a further investigation of this, and potential connections between the entanglement of scaling and the renormalized entanglement entropy [23] to future work. Furthermore, since our approach views RG as an operation on a QFT state, the RG monotones we find characterize a particular flow from the UV to the IR. An interesting question to explore is whether this quantity can be read off, directly from a CFT Hilbert space.

Acknowledgments

We are greatly indebted to Hong Liu for many valuable discussions on renormalization group flow. Also, we would like to thank Laurent Chaurette, Matthew Headrick, Petr Kravchuk, Juan Maldacena, Srivatsan Rajagopal and Matthew Roberts for informative discussions.

JHEP01(2019)219

A The entanglement of scaling is monotonic

We are interested in the derivative:

$$\lim_{\mu\to 0} \partial_R S(\rho_R \| \mathcal{E}(\rho_\mu)) \geq 0. \tag{A.1}$$

We start by proving that the operations, $\mathcal{E}$ and $\mathcal{N}$ commute: $\mathcal{N}(\mathcal{E}(\rho)) = \mathcal{E}(\mathcal{N}(\rho))$. Split the system in two parts: the part that is traced out A, and the remaining part B. The matrix elements of $\mathcal{E}(tr_A\rho)$ are

$$\int [D\psi]_A \, \langle \psi_A (f^{-1})^* \phi_B^+ | \rho | \psi_A (f^{-1})^* \phi_B^- \rangle. \tag{A.2}$$

After a change of variables this is equal to

$$\int [D(f^{-1})^*\psi]_A \, \langle (f^{-1})^* \psi_A (f^{-1})^* \phi_B^+ | \rho | (f^{-1})^* \psi_A (f^{-1})^* \phi_B^- \rangle.$$

which is nothing but $tr_A\mathcal{E}(\rho)$.

Relative entropy is monotonic under a partial trace: $\mathcal{N}_{R\to R-\delta R}$. We have

$$\begin{aligned} S\left(\rho_R \| \mathcal{E}(\rho_\mu)\right) \geq S\left(\mathcal{N}(\rho_R) \| \mathcal{N}\mathcal{E}(\rho_\mu)\right) &= S\left(\mathcal{N}(\rho_R) \| \mathcal{E}(\mathcal{N}(\rho_\mu))\right) \\ &= S(\rho_{R-\delta R} \| \mathcal{E}(\rho_\mu) + \mu\mathcal{E}(\delta\rho_\mu)) \end{aligned} \tag{A.3}$$

Taking the limit $\mu \to 0$ we establish that

$$\partial_R \mathcal{S}_{sc}(R) \geq 0. \tag{A.4}$$

References

[1] A.B. Zamolodchikov, *Irreversibility of the flux of the renormalization group in a 2D field theory*, *JETP Lett.* **43** (1986) 730 [*Pisma Zh. Eksp. Teor. Fiz.* **43** (1986) 565] [INSPIRE].

[2] J.L. Cardy, *Is there a c theorem in four-dimensions?*, *Phys. Lett.* **B 215** (1988) 749 [INSPIRE].

[3] Z. Komargodski and A. Schwimmer, *On renormalization group flows in four dimensions*, *JHEP* **12** (2011) 099 [arXiv:1107.3987] [INSPIRE].

[4] H. Casini and M. Huerta, *On the RG running of the entanglement entropy of a circle*, *Phys. Rev.* **D 85** (2012) 125016 [arXiv:1202.5650] [INSPIRE].

[5] H. Casini, E. Teste and G. Torroba, *Modular Hamiltonians on the null plane and the Markov property of the vacuum state*, *J. Phys.* **A 50** (2017) 364001 [arXiv:1703.10656] [INSPIRE].

[6] H. Casini, E. Testé and G. Torroba, *Markov property of the conformal field theory vacuum and the a theorem*, *Phys. Rev. Lett.* **118** (2017) 261602 [arXiv:1704.01870] [INSPIRE].

JHEP01(2019)219

[7] E.H. Lieb and M.B. Ruskai, *Proof of the strong subadditivity of quantum-mechanical entropy*, *J. Math. Phys.* **14** (1973) 1938 [INSPIRE].

[8] E.T. Jaynes, *Information theory and statistical mechanics*, *Phys. Rev.* **106** (1957) 620 [INSPIRE].

[9] P. Hayden, R. Jozsa, D. Petz and A. Winter, *Structure of states which satisfy strong subadditivity of quantum entropy with equality*, *Commun. Math. Phys.* **246** (2004) 359.

[10] O. Fawzi and R. Renner, *Quantum conditional mutual information and approximate Markov chains*, *Commun. Math. Phys.* **340** (2015) 575 [arXiv:1410.0664].

[11] M. Junge, R. Renner, D. Sutter, M.M. Wilde and A. Winter, *Universal recovery maps and approximate sufficiency of quantum relative entropy*, *Annales Henri Poincaré* **19** (2018) 2955 [arXiv:1509.07127] [INSPIRE].

[12] D. Petz, *Quantum information theory and quantum statistics*, Springer Science and Business Media, Germany (2007).

[13] D. Poulin and M.B. Hastings, *Markov entropy decomposition: a variational dual for quantum belief propagation*, *Phys. Rev. Lett.* **106** (2011) 080403.

[14] B. Czech, P. Hayden, N. Lashkari and B. Swingle, *The information theoretic interpretation of the length of a curve*, *JHEP* **06** (2015) 157 [arXiv:1410.1540] [INSPIRE].

[15] T. Faulkner, R.G. Leigh, O. Parrikar and H. Wang, *Modular Hamiltonians for deformed half-spaces and the averaged null energy condition*, *JHEP* **09** (2016) 038 [arXiv:1605.08072] [INSPIRE].

[16] T. Faulkner, R.G. Leigh and O. Parrikar, *Shape dependence of entanglement entropy in conformal field theories*, *JHEP* **04** (2016) 088 [arXiv:1511.05179] [INSPIRE].

[17] H. Casini, M. Huerta and R.C. Myers, *Towards a derivation of holographic entanglement entropy*, *JHEP* **05** (2011) 036 [arXiv:1102.0440] [INSPIRE].

[18] P. Calabrese and J.L. Cardy, *Entanglement entropy and quantum field theory*, *J. Stat. Mech.* **06** (2004) P06002 [hep-th/0405152] [INSPIRE].

[19] A.C. Wall, *A proof of the generalized second law for rapidly changing fields and arbitrary horizon slices*, *Phys. Rev.* **D 85** (2012) 104049 [*Erratum ibid.* **D 87** (2013) 069904] [arXiv:1105.3445] [INSPIRE].

[20] M.M. Roberts, *Time evolution of entanglement entropy from a pulse*, *JHEP* **12** (2012) 027 [arXiv:1204.1982] [INSPIRE].

[21] H. Casini, M. Huerta, R.C. Myers and A. Yale, *Mutual information and the F-theorem*, *JHEP* **10** (2015) 003 [arXiv:1506.06195] [INSPIRE].

[22] H. Casini, E. Teste and G. Torroba, *Relative entropy and the RG flow*, *JHEP* **03** (2017) 089 [arXiv:1611.00016] [INSPIRE].

[23] H. Liu and M. Mezei, *A refinement of entanglement entropy and the number of degrees of freedom*, *JHEP* **04** (2013) 162 [arXiv:1202.2070] [INSPIRE].

JHEP01(2019)219

Entanglement wedge cross sections require tripartite entanglement

Published for SISSA by Springer

Received: December 25, 2019
Accepted: April 10, 2020
Published: April 30, 2020

Entanglement wedge cross sections require tripartite entanglement

Chris Akers[a] **and Pratik Rath**[b,c]

[a] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.*
[b] *Center for Theoretical Physics and Department of Physics, University of California, Berkeley, CA 94720, U.S.A.*
[c] *Lawrence Berkeley National Laboratory, Berkeley, CA 94720, U.S.A.*

E-mail: cakers@mit.edu, pratik_rath@berkeley.edu

Abstract: We argue that holographic CFT states require a large amount of tripartite entanglement, in contrast to the conjecture that their entanglement is mostly bipartite. Our evidence is that this mostly-bipartite conjecture is in sharp conflict with two well-supported conjectures about the entanglement wedge cross section surface EW. If EW is related to either the CFT's reflected entropy or its entanglement of purification, then those quantities can differ from the mutual information at $\mathcal{O}(\frac{1}{G_N})$. We prove that this implies holographic CFT states must have $\mathcal{O}(\frac{1}{G_N})$ amounts of tripartite entanglement. This proof involves a new Fannes-type inequality for the reflected entropy, which itself has many interesting applications.

Keywords: AdS-CFT Correspondence, Gauge-gravity correspondence

ArXiv ePrint: 1911.07852

https://doi.org/10.1007/JHEP04(2020)208

JHEP04(2020)208

Contents

1 Introduction **1**

2 S_R conjecture vs bipartite entanglement **4**
- 2.1 Background 4
- 2.2 S_R of the bipartite entangled state 5
- 2.3 Small corrections 7
- 2.4 Tensor networks 10

3 E_P conjecture vs bipartite entanglement **13**

4 Discussion **15**

1 Introduction

We better understand quantum gravity every time we learn quantum information theoretic properties of holographic CFT states. This is the spirit of the "Geometry from Entanglement" slogan [1, 2], and it has been borne out in numerous discoveries. At the heart of these quantum information properties is the entanglement structure of the holographic CFT state. Know the structure explicitly, and you can in principle compute whatever quantum information property you want.

Hence it has been of great interest to probe this structure in any way tractable. Perhaps the most famous probe is a region's von Neumann entropy, whose bulk dual is simply the area divided by $4G_N$ of the minimal-area codimension-2 surface anchored to the boundary of the region [3, 4]. This is the Ryu-Takayanagi (RT) formula. It is well-known that the RT formula places strong constraints on the entanglement structure of the CFT state [5].

That said, the von Neumann entropy is a rather coarse measure of entanglement. It works well to quantify entanglement in a bipartite pure state, but doesn't capture all the information about entanglement structure for bipartite mixed states or multipartite states. Hence there is much less known about the multipartite structure of entanglement in holography, owing both to the fact that there have been fewer probes of it and that it is much harder to quantify (although there has been limited progress [6]).

It was in this context that a particularly powerful conjecture, which we call the "Mostly-Bipartite Conjecture" (MBC), was made by Cui et al. in [7]. We state this conjecture in detail now, as we understand it.

Mostly-bipartite conjecture of [7]. *Consider a state of a holographic CFT with a gravitational dual well-described by semiclassical gravity. Let $c \sim \frac{1}{G_N}$ represent its central charge. Given CFT subregions A,B, and C with Hilbert spaces that each admit the decomposition $H_X = H_{X_1} \otimes H_{X_2} \otimes H_{X_3}$, the quantum state is "close" to the form*

$$|\psi\rangle_{ABC} = U_A\, U_B\, U_C\, |\psi_1\rangle_{A_1\, B_1}\, |\psi_2\rangle_{A_2\, C_1}\, |\psi_3\rangle_{B_2\, C_2}\, |\widetilde{\psi}\rangle_{A_3\, B_3\, C_3} \tag{1.1}$$

JHEP04(2020)208

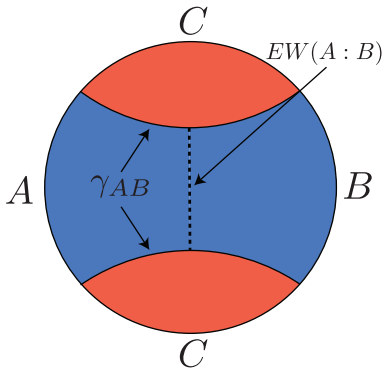


Figure 1. The entanglement wedge of boundary subregion AB is shaded blue, while the complementary entanglement wedge, corresponding to boundary subregion C, is shaded red. The RT surface is γ_{AB} (solid line), and the minimal cross section of the entanglement wedge is $EW(A:B)$ (dashed line).

in the $G_N \to 0$ *limit, where we demand that* $|\widetilde{\psi}\rangle_{A_3 B_3 C_3}$ *is 'small' in the sense that its entropies are subleading in* G_N*,*

$$S(A_3), S(B_3), S(C_3) \sim \mathcal{O}(1)\,, \tag{1.2}$$

while

$$S(A_1) = S(B_1) \approx \frac{I(A:B)}{2}\,, \tag{1.3}$$

$$S(A_2) = S(C_1) \approx \frac{I(A:C)}{2}\,, \tag{1.4}$$

$$S(B_2) = S(C_2) \approx \frac{I(B:C)}{2}\,, \tag{1.5}$$

where the "$\approx$*" symbol means at* $\mathcal{O}(\frac{1}{G_N})$*, and the mutual information is defined as* $I(A:B) \equiv S(A) + S(B) - S(AB)$*.*

We will refer to this conjectured state (1.1) as the "MBC state" from now on. We place quotes around "close" because it is not specified in what sense the states should be close. As we discuss in detail below, we will take this to mean close in natural distance measures usually applied to quantum states.

The motivation for this conjecture comes from the bit threads paradigm, in which Cui et al. found that an optimal bit thread configuration with the above bipartite structure exists. Moreover, this simple entanglement structure is realized by random stabilizer tensor networks (RSTNs), which are simple toy models of holography in which the RT formula is satisfied [8, 9].

Our goal is to argue that this entanglement structure is inconsistent with two other conjectured properties of AdS/CFT. Both of these other conjectures relate the so-called "minimal entanglement wedge cross section" $EW(A:B)$, of any two CFT subregions A and B, to information theoretic quantities of the CFT. We review these quantities in detail later, though see figure 1 for a quick visual. In the paper [10], the authors conjectured that

JHEP04(2020)208

$EW(A:B)$ equals one half a quantity called the reflected entropy, $S_R(A:B)$. The evidence for this conjecture is very strong, and we review it later. In the papers [11, 12], the authors conjectured that $EW(A:B)$ equals a quantity called the entanglement of purification, $E_P(A:B)$. There is also good evidence for this conjecture [13–15]. We shall refer to these as the S_R and E_P conjectures respectively.

Both S_R and E_P are more sensitive probes of multipartite entanglement than the von Neumann entropy is. It is this fact that places the S_R and E_P conjectures in tension with the MBC. Notably, our argument only works if either the S_R or E_P conjecture is true. This is because directly computing S_R and E_P is difficult, so we use their respective conjectures to compute them using the bulk.

The argument. In detail, our argument proceeds in two steps. First, we compute the reflected entropy and entanglement of purification of the state (1.1) and find that S_R equals the mutual information — and E_P half the mutual information — at leading order, $\mathcal{O}(\frac{1}{G_N})$. This is *not* true of holographic states, if either the S_R or E_P conjecture is correct. It is known that $2EW(A:B) - I(A:B)$ can be non-zero at $\mathcal{O}(\frac{1}{G_N})$, which implies $S_R - I$ and $2E_P - I$ should be non-zero at leading order as well. Therefore the MBC is in tension with the S_R and E_P conjectures.

That said, it is not obvious that this tension persists under small corrections to the MBC state. Indeed, it is conceivable that some sort of small correction to (1.1) could affect its S_R and E_P at $\mathcal{O}(\frac{1}{G_N})$ while *not* affecting other quantities, such as its von Neumann entropy, at that order. In that case, there would be no tension between these conjectures, because at any finite G_N the state would be of the MBC form up to subleading corrections and also have the correct S_R and E_P. Something like this is true for Renyi entropies, where exponentially small changes to a state can affect the Renyi entropy at $\mathcal{O}(\frac{1}{G_N})$ but only change the von Neumann entropy an exponentially small amount.

The second step in our argument is to prove that S_R and E_P are not sensitive to such small changes in the state. More precisely, we prove that S_R and E_P satisfy a Fannes-like continuity inequality so that when the trace distance $\frac{1}{2}||\rho - \sigma||_1$ between ρ and σ is ϵ, we have

$$|S_R(A:B)_\rho - S_R(A:B)_\sigma| \leq C_1 \sqrt{\epsilon} \log d\,, \tag{1.6}$$

$$|E_P(A:B)_\rho - E_P(A:B)_\sigma| \leq C_2 \sqrt{\epsilon} \log d\,, \tag{1.7}$$

where C_1, C_2 are $\mathcal{O}(1)$ constants and d is the dimension of ρ and σ. Moreover, we argue that $\epsilon < \mathcal{O}(1)$ if ρ is a holographic CFT state and σ is a state of the form eq. (1.1). (Otherwise, ρ would not take the MBC state form when $G_N \to 0$.) So, even though $\log d \sim \mathcal{O}(\frac{1}{G_N})$, the S_R and E_P of ρ is not different from that of σ at $\mathcal{O}(\frac{1}{G_N})$. Therefore, small corrections to eq. (1.1) that vanish as $G_N \to 0$ do not resolve the tension between these conjectures.

Why trace distance? Before proceeding, let us motivate why we use the trace distance to quantify small corrections. The trace distance is arguably the most natural distance measure between two quantum states. If two states are close in trace distance, then all observables computed using one will be close to those computed using the other, inlcuding

JHEP04(2020)208

the von Neumann entropy. Moreover, other distance measures (such as the fidelity) are quantitatively equivalent to trace distance. There are some quantities, like the relative entropy, that quantify the similarity of two states but are not technically distance measures. The relative entropy would work equally well for our purposes: if the relative entropy between two states is small, then their trace distance is small due to Pinsker's inequality.

That said, there are some senses in which two states can be "close" without being close in trace distance. For example, they can be "close" in the sense that some restricted class of observables has similar values. It is this sense in which, for instance, "random states" are close to "Perfect states." Perfect states are $2n$-partite states that are maximally entangled accross any bipartition, for n integer [16]. We define a random state by acting a Haar random unitary on a fiducial $2n$-partite state. Such random states are "close" to Perfect in the sense that they are nearly maximally entangled accross any bipartition. However, they are generally far from Perfect in trace distance.[1]

We choose not to consider "closeness" in this weaker sense because it is arguably against the spirit of the conjecture. Indeed, that the von Neumann entropies of holographic CFT states match those of the MBC state was the *motivation* for the MBC. The conjecture itself, as we understand it, is that the states are therefore close in some distance measure. Inferring this stronger claim about the state from the weaker matching of entropies is what makes the conjecture so valuable.

Organization. The paper is organized as follows. We define and analyze the S_R and E_P conjectures in section 2 and 3 respectively. Also in section 2, we discuss why RSTNs — which satisfy the RT formula — fail to satisfy the S_R conjecture, which naively seems like a simple application of RT. We briefly touch on tensor networks in section 3 as well. Finally, we conclude with some discussion and future directions in section 4.

Notation. We will use the notation $S_R(A:B)$, $E_P(A:B)$ and $I(A:B)$ to denote the reflected entropy, entanglement of purification and mutual information relevant for the partition of the state about subregions A and B. However, in other situations where the partition is understood and we would like to make explicit the state in which these quantities are being evaluated, we shall use the notation $S_R(\rho_{AB})$, $E_P(\rho_{AB})$ and $I(\rho_{AB})$ interchangeably with the above notation.

2 S_R conjecture vs bipartite entanglement

2.1 Background

We now define the reflected entropy $S_R(A:B)$. Consider a density matrix ρ_{AB} on the Hilbert space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$. One can define its "canonical purification" in a way analogous to the relationship between the thermal density matrix and the thermofield double state [10]. There exists a natural mapping between the space of linear operators acting on

[1] This can be seen from a simple counting argument: there are far fewer Perfect states than the total number of states. In the limit that the Hilbert space dimension goes to infinity, the average distance between any given state and the nearest Perfect state tends to zero.

JHEP04(2020)208

a $\mathcal{H}$ and the space of states on a doubled Hilbert space $\mathcal{H} \otimes \mathcal{H}' = \mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_{A'} \otimes \mathcal{H}_{B'}$. This mapping is sometimes labelled the channel-state duality. The inner product on this doubled Hilbert space is defined by

$$\langle \rho | \sigma \rangle_{ABA'B'} = \mathrm{tr}_{AB}(\rho^\dagger \sigma)\,. \tag{2.1}$$

Thus, the operator $\sqrt{\rho_{AB}}$ can be mapped to a state $|\sqrt{\rho_{AB}}\rangle_{ABA'B'}$, which is named the canonical purification of ρ_{AB} (and is also known as the GNS state). This state easily can be checked to reduce to the original density matrix ρ_{AB} upon tracing out the subregions A' and B'. Given the above setup, then

Definition 2.1. *The* reflected entropy $S_R(A:B)$ *is defined as*

$$S_R(A:B) = S(AA')_{\sqrt{\rho_{AB}}} = S(BB')_{\sqrt{\rho_{AB}}}\,, \tag{2.2}$$

where $S(AA')_{\sqrt{\rho_{AB}}}$ *is the von Neumann entropy of the reduced density matrix on the subregion* AA' *in the state* $|\sqrt{\rho_{AB}}\rangle$.

In [10], it was conjectured that in AdS/CFT,

$$2EW(A:B) = S_R(A:B)\,, \tag{2.3}$$

where $EW(A:B)$ is the area of the "entanglement wedge cross-section," i.e. the minimal-area surface that divides the entanglement wedge of AB into two halves, one homologous to A and the other to B. This conjecture is intuitive: the reduced density matrix of AB is unchanged, and $A'B'$ has the same reduced density matrix. One can solve the equations of motion inwards from this data local to the boundary to conclude that a viable bulk solution is the one that is simply two copies of the AB entanglement wedge glued together across the extremal surface that bounds it. (The subtleties of gluing across this extremal surface were discussed in [17].) Applying the RT formula to the AA' region of this doubled bulk implies that $S(AA')_{\sqrt{\rho_{AB}}}$ equals the area of a minimal surface dividing AA' from BB'. The symmetry between the entanglement wedges of AB and $A'B'$ implies that this minimal surface has area $2EW$.[2]

2.2 S_R of the bipartite entangled state

We now compute the reflected entropy in the MBC state eq. (1.1) and show that it approximately equals the mutual information,

$$S_R(A:B) \approx I(A:B)\,. \tag{2.4}$$

This, we will argue, is incompatible with AdS/CFT. Two properties of the reflected entropy will be useful to us. First, it is an additive quantity under tensor products:

$$S_R(\rho_1 \otimes \rho_2) = S_R(\rho_1) + S_R(\rho_2)\,. \tag{2.5}$$

[2]Evidence for the conjecture in a time-dependent situation was provided in [18, 19].

JHEP04(2020)208

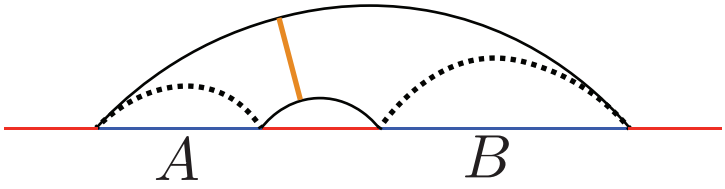


Figure 2. Subregion AB at the threshold of a mutual information phase transition. There are two competing RT surfaces, denoted by solid and dashed black lines. The area of the dashed lines is equal to the area of the solid lines. $EW(A:B)$ before the transition is denoted by a solid orange line, while it vanishes after the transition.

This is because the canonical purification of a tensor product density matrix $\rho_1 \otimes \rho_2$ is given by the tensor product state $|\sqrt{\rho_1}\rangle \otimes |\sqrt{\rho_2}\rangle$. Second, the reflected entropy is invariant under unitaries local to A or B, since this is equivalent to local unitaries on A, A', B and B' in the purified state. Hence the reflected entropy of the MBC state is the same as for the state

$$U_A^\dagger U_B^\dagger \rho_{AB} U_A U_B = \rho_{A_1\,B_1} \otimes \rho_{A_2} \otimes \rho_{B_2} \otimes \rho_{A_3 B_3}\,, \tag{2.6}$$

where e.g. $\rho_{A_2} = \mathrm{tr}_{C_1} |\psi_2\rangle \langle\psi_2|_{A_2 C_1}$. Thus, the calculation of S_R splits into an individual calculation for each factor. First consider $\rho_{A_1\,B_1} = |\psi_1\rangle \langle\psi_1|_{A_1 B_1}$. The canonical purification is simply a product state of two copies of $|\psi_1\rangle$, and therefore

$$S_R\left(\rho_{A_1 B_1}\right) = 2S(\rho_{A_1}) = I(A_1:B_1)_{\rho_{A_1 B_1}} \approx I(\rho_{AB})\,. \tag{2.7}$$

Because the state ρ_{A_2} only has support on A, its canonical purification is given by an entangled state shared between A and A' while B and B' remain trivial. The same argument can be applied to ρ_{B_2} as well. Therefore their reflected entropies vanish,

$$S_R\left(\rho_{A_2}\right) = 0 \quad \text{and} \quad S_R(\rho_{B_2}) = 0\,. \tag{2.8}$$

Although we have not specified any details of the state $|\tilde{\psi}\rangle_{A_3\,B_3\,C_3}$, we can use the general inequality

$$S_R(\rho_{A_3\,B_3}) \leq 2\,\min\{S(\rho_{A_3}), S(\rho_{B_3})\} = O(1) \tag{2.9}$$

to put an upper bound on the contribution to S_R from $\rho_{A_3 B_3}$. It is a positive $\mathcal{O}(1)$ number, at most. Putting everything together, we find that the reflected entropy equals

$$\begin{aligned} S_R(\rho_{AB}) &= S_R(\rho_{A_1\,B_1}) + S_R(\rho_{A_2}) + S_R(\rho_{B_2}) + S_R(\rho_{A_3\,B_3}) \\ &= I(\rho_{AB}) + O(1)\,. \end{aligned} \tag{2.10}$$

Hence in the $G_N \to 0$ limit, $S_R(A:B) = I(A:B)$ for the MBC state.

AdS/CFT conflict. We now argue that this is in conflict with $S_R(A:B) = 2EW(A:B)$ in AdS/CFT. The idea is that $EW(A:B)$ can be larger than $I(A:B)$ at $\mathcal{O}(\frac{1}{G_N})$. This is true in many generic cases, but we now provide a sharp example in which this is especially clear, from [20].

JHEP04(2020)208

Consider the setup in figure 2. As one varies the distance between subregions A and B of a fixed size, one encounters a phase transition in the RT surfaces. At the phase transition, both $I(A:B)$ and $EW(A:B)$ vanish. However, at slightly shorter separations the two are quite different. While the mutual information continuously shrinks to zero as the separation is increased, the cross-section remains $\mathcal{O}(\frac{1}{G_N})$ until exactly at the phase transition, where it discontinuously jumps to zero. Therefore, given $S_R(A:B) = 2EW(A:B)$, we must conclude that the MBC state is incompatible with AdS/CFT.

2.3 Small corrections

So far, we have not ruled out that the S_R conjecture is consistent with the MBC state *with small corrections.* One might imagine that the reflected entropy, being non-linear in the state, could receive large corrections from terms that are subleading in G_N to those in eq. (1.1).[3] Then there would be no tension between the S_R conjecture and MBC: For any finite G_N, the holographic CFT state could take the form of the MBC state up to subleading terms, but its reflected entropy could be different at $\mathcal{O}(\frac{1}{G_N})$. For comparison, this is how Renyi entropies work. Renyi entropies are also non-linear in the state, and can change at $\mathcal{O}(\frac{1}{G_N})$ under non-perturbatively small changes to the state.

We quantify corrections to the state in terms of the natural distance measure, trace distance, defined as

$$T(\rho,\sigma) = \frac{1}{2}||\rho-\sigma||_1\,, \tag{2.11}$$

where ρ, σ are two density matrices, and $||A||_1 = \mathrm{tr}(\sqrt{A^\dagger A})$ is the Schatten 1-norm or L_1 norm. It can take values $T(\rho,\sigma)\in[0,1]$, and when the trace distance is close to 0 then all observables are close between the states. If the trace distance is exactly zero, then the two states are identically equal. If two states admit a G_N expansion, like $\rho = \rho_0 + G_N\rho_1 + \mathcal{O}(G_N^2)$, then the trace distance between them does as well:

$$T(\rho,\sigma) = T_0(\rho,\sigma) + G_N T_1(\rho,\sigma) + \mathcal{O}(G_N^2)\,. \tag{2.12}$$

We say that two states are the same at leading order if $T_0 = 0$, i.e. $T(\rho,\sigma)\sim\mathcal{O}(G_N)$.[4] For our purposes, we could equally-well use other distance measures between states, such as the fidelity, or similarity measures like the relative entropy.

We interpret the MBC as the statement the trace distance vanishes at leading order in G_N between a holographic CFT state ρ and some state σ of the form eq. (1.1). This is for two reasons. First, as stated above, so that ρ and σ become the same in the $G_N\to 0$ limit. Second, because this would give a satisfactory reason for the von Neumann entropies to match at leading order (even at finite G_N). (After all, this was essentially the motivation for the conjecture in the first place!) This is due to Fannes inequality [21], which states

$$|S(\rho)-S(\sigma)| \le 2T(\rho,\sigma)\log d - 2T(\rho,\sigma)\log(2T(\rho,\sigma))\,, \tag{2.13}$$

[3] We would like to thank Matt Headrick for discussions related to this.

[4] In fact, for the purpose of our analysis $T(\rho,\sigma)\sim\mathcal{O}(G_N^a)$ with any $a>0$ works.

JHEP04(2020)208

where d is the dimension of ρ and σ. For holographic CFTs, $\log d \sim \mathcal{O}(\frac{1}{G_N})$, and thus if $T(\rho,\sigma) \lesssim \mathcal{O}(G_N)$, the von Neumann entropies will be guaranteed to match at $\mathcal{O}(\frac{1}{G_N})$.

So, we are interested in whether the reflected entropy can differ at $\mathcal{O}(\frac{1}{G_N})$ between the MBC state σ and a holographic CFT state ρ that differs from it only at $\mathcal{O}(G_N)$ and higher,

$$T(\rho,\sigma) \sim \mathcal{O}(G_N)\,. \tag{2.14}$$

We now prove this is, in fact, not possible; the reflected entropy satisfies a continuity inequality similar to Fannes inequality for the von Neumann entropy.

Theorem 2.1 (Continuity of the Reflected Entropy)**.** *Given two density matrices ρ_{AB} and σ_{AB} defined on a Hilbert space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$ of dimension $d = d_A\, d_B$, such that $T_{AB} = T(\rho_{AB},\sigma_{AB}) \leq \epsilon$, then*

$$|S_R(\rho_{AB}) - S_R(\sigma_{AB})| \leq 4\sqrt{2T_{AB}}\log(\min\{d_A,d_B\}) - 2\sqrt{2T_{AB}}\log(T_{AB})$$

for $\epsilon \leq \frac{1}{8e^2}$.

Proof. In order to prove the above statement, we first consider the fidelity between the respective purified states $|\sqrt{\rho_{AB}}\rangle_{ABA'B'}$ and $|\sqrt{\sigma_{AB}}\rangle_{ABA'B'}$, which is given by

$$F_{ABA'B'} = |\,\langle\sqrt{\rho_{AB}}|\sqrt{\sigma_{AB}}\rangle\,|\,. \tag{2.15}$$

The inner product on the canonically purified states can equivalently be computed using the original density matrices by using eq. (2.1),

$$\langle\sqrt{\rho_{AB}}|\sqrt{\sigma_{AB}}\rangle = \mathrm{tr}(\sqrt{\rho_{AB}}\sqrt{\sigma_{AB}}) \tag{2.16}$$

$$= Q_{1/2}(\rho_{AB},\sigma_{AB}), \tag{2.17}$$

where $Q_{1/2}(\rho_{AB},\sigma_{AB})$ is defined by the above equation and is the non-commutative generalization of the Bhattacharya coefficient.[5] Now we can use the inequality [22]

$$Q_{1/2}(\rho_{AB},\sigma_{AB}) \geq 1 - T_{AB} \tag{2.18}$$

$$\implies F_{ABA'B'} = Q_{1/2}(\rho_{AB},\sigma_{AB}) \geq 1 - T_{AB}\,. \tag{2.19}$$

This is essentially equivalent to the well known Powers-Stormer inequality. Upon tracing out B and B', the fidelity monotonically increases giving us

$$F_{AA'} \geq F_{ABA'B'} \geq 1 - T_{AB}\,. \tag{2.20}$$

Now, we can use another well-known inequality relating fidelity to trace distance [22], giving us

$$T(\rho_{AA'},\sigma_{AA'}) \leq \sqrt{1 - F_{AA'}^2} \leq \sqrt{2T_{AB}}\,, \tag{2.21}$$

[5] Note that $Q_{1/2}$ is a real quantity, which can be proven using cyclicity of trace and the fact that density matrices are Hermitian.

JHEP04(2020)208

where e.g., $\rho_{AA'}$ is the density matrix obtained by tracing out BB' from the purified state $|\sqrt{\rho_{AB}}\rangle$. The second inequality in eq. (2.21) follows from eq. (2.20). Thus, starting from ρ and σ being ϵ-close in trace distance on subregion AB, we have shown that their canonical purifications are $\sqrt{\epsilon}$-close in trace distance on subregion AA'. Finally, we use Fannes inequality [21] to show that

$$\begin{aligned} |S_R(A:B)_\rho - S_R(A:B)_\sigma| &= |S(\rho_{AA'}) - S(\sigma_{AA'})| \\ &\leq 2T_{AA'}\log(d_{AA'}) - 2T_{AA'}\log(2T_{AA'}) \\ &\leq 4\sqrt{2T_{AB}}\log(d_A) - 2\sqrt{2T_{AB}}\log(T_{AB}), \end{aligned} \tag{2.22}$$

where $T_{AA'} = T(\rho_{AA'}, \sigma_{AA'})$.[6] This inequality holds for $T_{AA'} \leq \frac{1}{2e}$, which is ensured by the bound $\epsilon \leq \frac{1}{8e^2}$. The entire analysis above was perfectly symmetric between A and B, and from eq. (2.2) we also have

$$|S_R(A:B)_\rho - S_R(A:B)_\sigma| \leq 4\sqrt{2T_{AB}}\log(d_B) - 2\sqrt{2T_{AB}}\log(T_{AB}). \tag{2.23}$$

Thus, combining eq. (2.22) and eq. (2.23), we get the strengthened inequality

$$|S_R(A:B)_\rho - S_R(A:B)_\sigma| \leq 4\sqrt{2T_{AB}}\log(\min\{d_A, d_B\}) - 2\sqrt{2T_{AB}}\log(T_{AB})\,, \tag{2.24}$$

which proves Theorem 2.1. □

Note that it was crucial that we considered the *canonical* purification in order for e.g. $|S(\rho_{AA'}) - S(\sigma_{AA'})|$ to have such a bound. An arbitrary purification on $ABA'B'$ can be arbitrarily far in trace distance. For example, different Bell pairs purify a maximally mixed density matrix and have trace distance 1. The canonical purification ensures this redundancy in basis of purification doesn't play a role here.

We also emphasize that we have not found any examples where the inequality in Theorem 2.1 is saturated, despite the fact that it is easy to saturate all the individual inequalities required in proving it. Our preliminary numerical analysis suggests that $|S_R(\rho) - S_R(\sigma)| \sim O(\epsilon)$ in all the examples that we tested, instead of the $O(\sqrt{\epsilon})$ allowed by Theorem 2.1. This leaves open the possibility that a tighter bound exists. We haven't pursued a systematic numerical analysis of the above, but it would be interesting to probe this question in future.

Implication for AdS/CFT. Theorem 2.1 renders it impossible for two states ρ_{AB}, σ_{AB} to have reflected entropy different at $\mathcal{O}(\frac{1}{G_N})$ unless $\sqrt{T_{AB}}\log d_{AB}$ is also $\mathcal{O}(\frac{1}{G_N})$. In a holographic CFT, $\log d_{AB} \sim \mathcal{O}(\frac{1}{G_N})$. So, the trace distance would need to be non-zero at leading order, $T_{AB} \sim \mathcal{O}(1)$.

However, this is not consistent with the MBC. Suppose σ_{ABC} represents the density matrix corresponding to the MBC state, and ρ_{ABC} represents the actual density matrix of a holographic CFT. As we argued above, the MBC requires they should be close in the sense that $T_{ABC} \equiv T(\rho_{ABC}, \sigma_{ABC}) \sim \mathcal{O}(G_N)$. Trace distances decrease under tracing out

[6] This result can be further tightened by using the Audenaart version of the inequality [23].

JHEP04(2020)208

subregions, so $T_{AB} \leq T_{ABC} \sim \mathcal{O}(G_N)$. Therefore, T_{AB} is too small for σ and ρ to have different reflected entropy at $\mathcal{O}(\frac{1}{G_N})$.

Said differently, Theorem 2.1 states that if T_{ABC} is indeed $\mathcal{O}(G_N)$, then

$$|S_R(\rho_{AB}) - S_R(\sigma_{AB}))| = |2EW(A:B) - I(A:B)| \lesssim \mathcal{O}\left(\frac{1}{\sqrt{G_N}}\right) , \tag{2.25}$$

where we have used the S_R conjecture in the equality and Theorem 2.1 in the inequality. This contradicts the fact that there exist examples in AdS/CFT where $|2EW(A:B) - I(A:B)| \sim \mathcal{O}(\frac{1}{G_N})$, e.g. the situation in figure 2. Thus, we see that even small corrections to the MBC state are incapable of making it compatible with the S_R conjecture.

2.4 Tensor networks

We now resolve a conundrum that our results seem to create in tensor networks. Tensor networks have provided good toy models of holography, illustrating properties such as subregion duality and the RT formula. In particular, a network made of perfect tensors can be shown to satisfy the RT formula under certain reasonable assumptions [16]. Much more generally, it was shown that networks made from Haar random tensors also satisfy the RT formula [8].

It was also emphasized in [8] that Haar randomness was overkill, and the RT formula followed simply from choosing random tensors from a 2-design ensemble, i.e. one that agrees with the first two moments of the Haar measure. A particularly nice choice of 2-design ensemble is provided by stabilizer tensors of dimension $D = p^N$ in the limit of large N, where p is a prime number. Such random stabilizer tensor networks (RSTN) were further studied in [9], where it was proven that their states always take the form

$$|\psi\rangle_{ABC} = U_A^\dagger U_B^\dagger U_C^\dagger \, |\phi^+\rangle_{A_1 B_1}^{\otimes n_1} \, |\phi^+\rangle_{A_2 C_1}^{\otimes n_2} \, |\phi^+\rangle_{B_2 C_2}^{\otimes n_3} \, |\mathrm{GHZ}\rangle_{A_3 B_3 C_3}^{\otimes n_g} \tag{2.26}$$

where $|\phi^+\rangle$ denotes a p-dimensional Bell pair shared between the two parties, e.g.

$$|\phi^+\rangle_{A_1 B_1} \equiv \frac{1}{\sqrt{p}} \sum_{i=0}^{p-1} |i\rangle_{A_1} \, |i\rangle_{B_1} \ , \tag{2.27}$$

and $|\mathrm{GHZ}\rangle$ denotes a shared p-dimensional GHZ state,

$$|\mathrm{GHZ}\rangle_{A_3 B_3 C_3} = \frac{1}{\sqrt{p}} \sum_{i=0}^{p-1} |i\rangle_{A_3} \, |i\rangle_{B_3} \, |i\rangle_{C_3} \ . \tag{2.28}$$

Neither of these states scale with N; they are elementary units of entanglement. The exponents, however, can indeed have N-dependence. That N-dependence was discovered in [9], where it was shown that in the large N limit, n_1, n_2 and n_3 grow linearly with N, whereas n_g remains $\mathcal{O}(1)$. Note that N here is analogous to $\frac{1}{G_N}$ in AdS/CFT.

This is exactly an MBC state like that in eq. (1.1). Our result in section 2 shows that this is incompatible with the conjecture $S_R = 2EW$. This is startling at first: the S_R conjecture was motivated by the RT formula, which RSTN satisfy. So, naively, we would expect RSTN to satisfy $S_R = 2EW$.

JHEP04(2020)208

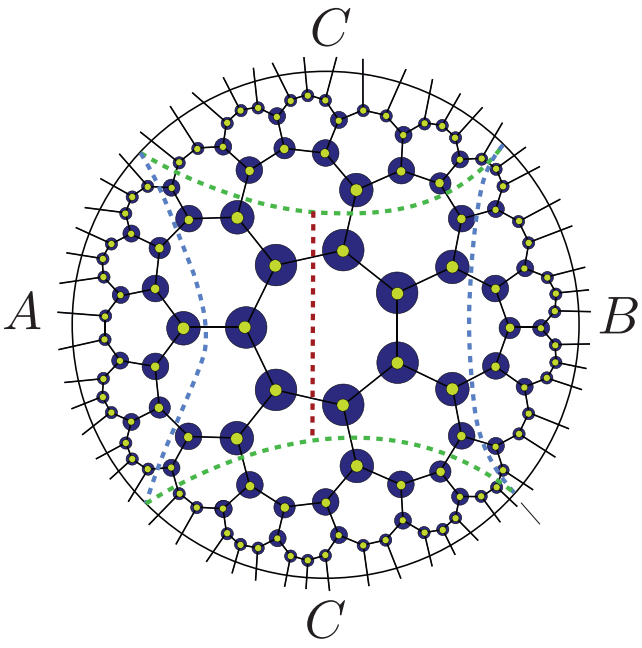


Figure 3. A random stabilizer tensor network with subregion AB in the connected phase. The green dotted line represents the RT surface for subregion AB, while the yellow dotted lines represent the RT surface of A and B respectively. The red dotted line represents $EW(A:B)$.

We now compute $S_R(A:B)$ in RSTN to explain why they, in fact, do not. The upshot will be that while the canonical purification of a state ρ_{AB} is indeed given by a doubled version of its entanglement wedge (just like in AdS/CFT), the doubled entanglement wedge network does not itself satisfy RT in the naive way!

Consider the tensor network in figure 3. In order to restrict to ρ_{AB}, we can use the fact that there is an isometry from the boundary legs of subregion C to the in-plane legs cut by the RT surface of subregion AB. This gives us an effective tensor network restricted to the entanglement wedge of AB. In order to compute the density matrix ρ_{AB}, we can glue together two copies of this tensor network as in figure 4. The density matrix ρ_{AB} has a flat entanglement spectrum as can be seen from eq. (2.26). Thus, it can be shown that the operator $\sqrt{\rho_{AB}}$, and hence the canonically purified state $|\sqrt{\rho_{AB}}\rangle_{ABA'B'}$ is represented by the same doubled tensor network TN' depicted in figure 4 up to normalization.

TN' geometrically resembles the bulk saddle geometry obtained in the holographic construction discussed in [10]. If TN' were to satisfy the RT formula, one would indeed be led to the claim that the entropy of subregion AA' is computed by the minimal cross section in this effective tensor network. The RT surface in TN' is indeed just twice the original entanglement wedge cross section, and thus, we would have the conjectured result, $S_R(A:B) = 2EW(A:B)$.

However, this naive argument doesn't carry through because TN' has certain special properties that distinguish it from a completely random stabilizer tensor network. Importantly, the set of tensors used in Copy 2 in TN' are precisely correlated with the tensors in Copy 1. E.g., in figure 4, one can see $T_1^\dagger$ and T_1 placed at equivalent positions in ei-

JHEP04(2020)208

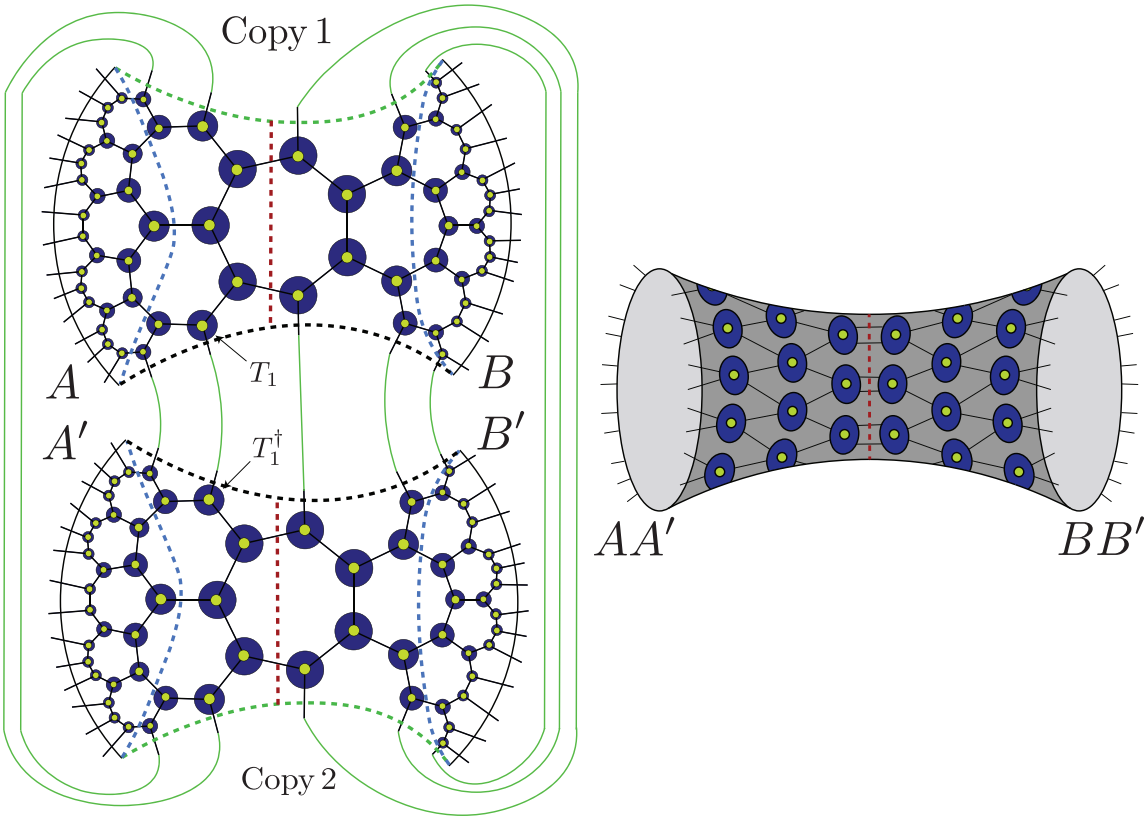


Figure 4. (Left): A reduced tensor network corresponding to the entanglement wedge of AB is obtained by using the isometry from the boundary legs of subregion C to the legs at the RT surface (denoted black and green dotted lines). Two copies of this RSTN glued as shown prepare the canonically purified state. We call this doubled network TN'. (Right): Geometrically, this resembles the AdS/CFT construction discussed in [10, 17, 24]. If the RT formula holds, then $S_R(A:B) = 2EW(A:B)$.

ther copy. The derivation of the RT formula depended on having completely uncorrelated tensors on both copies of the TN.

That this correlation spoils the RT formula is made manifest by the form of the state $|\psi\rangle_{ABC}$ in eq. (2.26). After applying the local unitaries, which depend sensitively on the choice of tensors in the network, one gets a drastically simplified network as seen in figure 5. The canonical purification then takes a simple form, and computing $S(AA')$ in this simple network gives us

$$S_R(A:B) = 2n_1 \log p = I(A:B)\,. \tag{2.29}$$

We see that RSTN do not satisfy $S_R = 2EW$ because having correlated tensors precludes the application of the RT formula.

Indeed, the RT formula in the original RSTN only required the tensors be 2-designs. We expect that having the tensors agree with even higher moments of the Haar measure is sufficient for the network to continue to satisfy the RT formula, even when the network is built out of many copies of itself. If true, then the random tensor networks of [8] should

JHEP04(2020)208

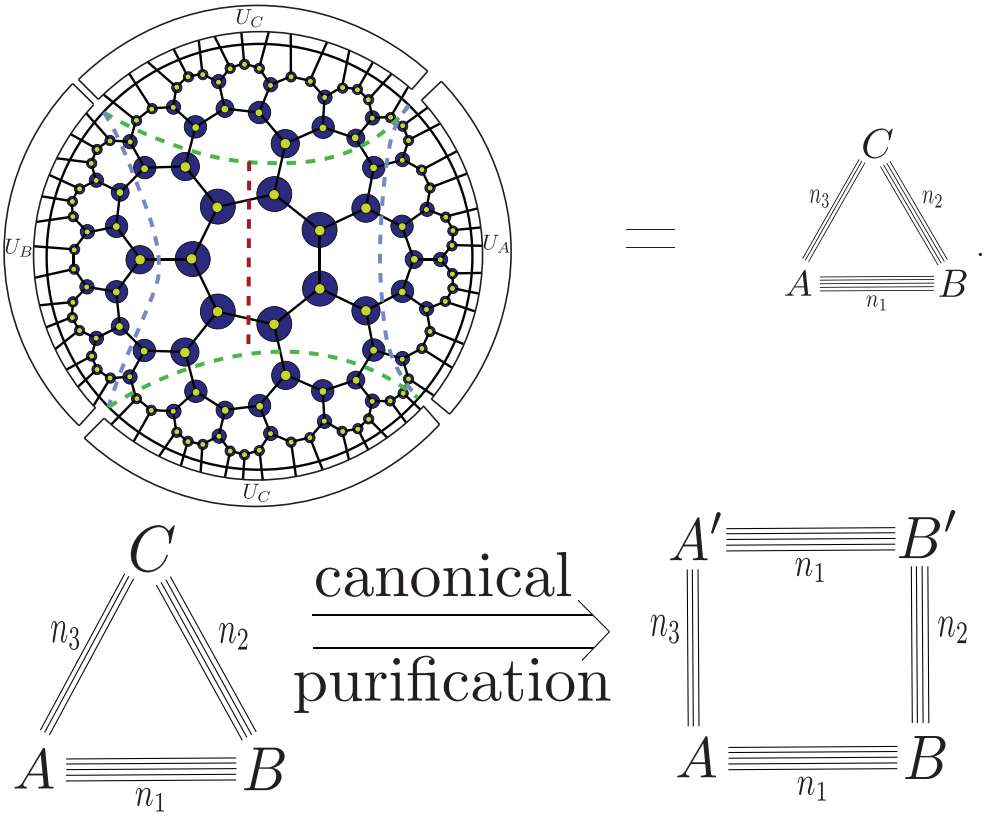


Figure 5. After applying local unitaries, the RSTN drastically simplifies to a combination of Bell pairs shared by the three parties. The Bell pairs then lead to a simple canonically purified state.

satisfy the S_R conjecture, and highly random tensors — rather than e.g. 2-designs — would be better models of holography. This is the subject of ongoing work [25].

3 E_P conjecture vs bipartite entanglement

There is a tension between the E_P conjecture and the MBC that is qualitatively the same as that between the S_R conjecture and the MBC. Given a density matrix ρ_{AB}, one can define its entanglement of purification as [26]

$$E_P(A:B) = \min_{|\psi\rangle} S(AA')\,, \tag{3.1}$$

where the minimization is over all states $|\psi\rangle_{ABA'B'}$ that are pure and consistent with the reduced density matrix ρ_{AB}. In [11, 12], it was conjectured that in AdS/CFT

$$E_P(A:B) = EW(A:B). \tag{3.2}$$

This conjecture was motivated by the surface-state correspondence, wherein similar to tensor networks, a holographic state can be defined on any convex surface in the bulk [27–30]. Further, since the minimization over all possible purifications is a computationally

JHEP04(2020)208

intractable problem, it was assumed that minimizing over geometric purifications was sufficient (for discussions of this point, see [31]). This conjecture, along with its multipartite generalizations, has received a lot of attention recently, although proofs or related computations have generally required various strong assumptions [13–15, 32–39].

To argue that the E_P conjecture is incompatible with the MBC, we review results that are essentially known in the literature. This distinguishes this argument from the one in section 2, which involved our Theorem 2.1 that was completely new.

In order to compute $E_P(A:B)$ in the MBC state, we first note that E_P is a subadditive quantity under tensor products [40]. In fact, additivity holds for pure states, $\rho_{AB} = |\psi\rangle_{AB}\langle\psi|_{AB}$, and completely decoupled states, $\rho_{AB} = \rho_A \otimes \rho_B$, but not in general [41]. Using this property, we find for the MBC state

$$E_P(\rho_{AB}) \leq E_P(\rho_{A_1\,B_1}) + E_P(\rho_{A_2}) + E_P(\rho_{B_2}) + E_P(\rho_{A_3\,B_3}). \tag{3.3}$$

The first term on the right hand side gives $E_P(\rho_{A_1B_1}) = S(\rho_{A_1}) = \frac{1}{2}I(A_1:B_1)$, because $\rho_{A_1B_1}$ is a pure state. The second and third terms involve only one of either A or B and thus give $E_P(\rho_{A_2}) = E_P(\rho_{B_2}) = 0$. The fourth term can be bounded using the known inequalities for E_P to obtain

$$0 \leq E_P(\rho_{A_3\,B_3}) \leq 2\min\{S(\rho_{A_3}), S(\rho_{B_3})\}, \tag{3.4}$$

and thus, $E_P(\rho_{A_3\,B_3})$ is an $O(1)$ positive quantity. Putting these results together and using known inequalities, we find that

$$\frac{1}{2}I(A:B) \leq E_P(\rho_{AB}) \leq \frac{1}{2}I(A:B) + O(1). \tag{3.5}$$

Thus, for $G_N \to 0$, we obtain $E_P(A:B) \approx \frac{1}{2}I(A:B)$, where " $\approx$" denotes matching at $\mathcal{O}(\frac{1}{G_N})$. Similar to the result in section 2.2, we find that the MBC state is incompatible with the E_P conjecture.

Small corrections. One might again worry that small corrections to the MBC state might make it compatible with the E_P conjecture. However, this too can be ruled out by the following theorem.

Theorem 3.1 (Continuity of the Entanglement of Purification)**.** *Given two density matrices ρ_{AB} and σ_{AB} defined on a Hilbert space $\mathcal{H} = \mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B}$ of dimension $d = d_A\, d_B$, such that $T_{AB} = T(\rho_{AB}, \sigma_{AB}) \leq \epsilon$, then*

$$|E_P(\rho_{AB}) - E_P(\sigma_{AB})| \leq 40\sqrt{T_{AB}}\log(d) - 4\sqrt{T_{AB}}\log(4\sqrt{T_{AB}})$$

for $\epsilon \leq \frac{1}{4e^2}$.

Proof. This proof essentially follows from Theorem 1 of [26], where it was shown that

$$|E_P(\rho_{AB}) - E_P(\sigma_{AB})| \leq 20D(\rho_{AB}, \sigma_{AB})\log(d) - D(\rho_{AB}, \sigma_{AB})\log(D(\rho_{AB}, \sigma_{AB})) \tag{3.6}$$

where $D(\rho_{AB}, \sigma_{AB}) = 2\sqrt{1 - F_{AB}}$ is the Bures distance. Using the inequality

$$1 - T_{AB} \leq F_{AB} \implies D(\rho_{AB}, \sigma_{AB}) \leq 2\sqrt{T_{AB}}, \tag{3.7}$$

we obtain the desired result. □

JHEP04(2020)208

Using Theorem 3.1, we conclude that a slightly-corrected MBC state is still incompatible with the E_P conjecture, by a similar argument to the one made in section 2.3.

Tensor networks. E_P is a difficult quantity to compute in general, and hence it is much harder to understand the tensor network story analogous to that in section 2.4. However, in the case of RSTNs, the simplified network (obtained by applying local unitaries, as in figure 5) has an E_P that can easily be calculated to give $\frac{1}{2}I(A:B)$ at leading order in G_N.

It is important to note that the E_P conjecture was originally motivated by restricting to geometric purifications and computing the optimal RT surface anchored to the entanglement wedge. An important insight we gain here is that non-geometric tensor networks like the simplified network were crucial for the minimization in computing E_P, at least for RSTNs. It would be interesting to understand if this is more generally true [31].

4 Discussion

We have provided two pieces of evidence that suggest that holographic states require a large amount of tripartite entanglement: Having little tripartite entanglement is inconsistent with both the strongly-supported conjectures that $S_R = 2EW$ and $E_P = EW$. We now focus on some of the caveats, implications, and interesting future directions stemming from this work.

Trace distance. We have demonstrated that holographic CFT states cannot be close in trace distance to the MBC state. It is still possible that they are "close" in another sense. Being close in trace distance is a strong criterion that ensures closeness in all observable quantities and is a standard measure of similarity of states in quantum information. If we allow weaker conditions of closeness on the state, such as closeness in a restricted class of observable quantities, it might be possible to make the MBC state consistent with the S_R and E_P conjectures. However, we do not see any evidence for other quantities that may be reproduced by assuming an MBC state, and in particular, measures of multipartite entanglement are in conflict with the conjectured state. It would be interesting to see if other weaker forms of closeness can lead to a version of the MBC that is both useful and compatible with the other two conjectures.

Limitation on tensor networks. This analysis also illuminates limitations of tensor networks as toy models of holography. Since the von Neumann entropy is a reasonably coarse grained quantity, even 2-design tensor networks such as random stabilizer tensor networks were able to reproduce the RT formula. However, stabilizers are a very special class of tensors, and are generically far in trace distance from Haar random tensors (owing to the fact that there are many more Haar random tensors than stabilizers). Hence, properties from any such tensor networks should be considered carefully, because they may not agree with actual holographic answers.

In fact, specific tensor network models have previously been used to model "mostly bipartite" entanglement that arises in certain regions of moduli space of multiboundary

JHEP04(2020)208

wormholes [42, 43]. It would be interesting to explore whether more refined tensor network models can capture the right form of multipartite entanglement employed by holographic states.

It is interesting to note that the tensor network in [44, 45] is close in trace distance to the holographic state, by construction. Certain classes of their tensor networks require the $E_P = EW$ conjecture, so it would be interesting to repeat the above analysis in their case.

Entanglement measures. As we saw in our analysis in section 2, the reflected entropy $S_R(A:B)$ is a much more fine-grained entanglement measure than individual entanglement entropies, for mixed density matrices. This quantity is very naturally motivated from holography and hasn't yet been studied in the quantum information literature. In this sense, it is similar to the refined Renyi entropies which is also a very natural quantity in holography, but hasn't been analyzed in quantum information [46–49]. It would be interesting to understand its properties and generic behaviour in quantum systems.

There is, in fact, a zoo of quantities that measure multipartite entanglement and there is not a clear understanding of a canonically best choice. Owing to this fact, there have been many proposals in holography for such quantities including, among many others, the entanglement negativity and odd entropy [20, 50–52]. Similarly, higher party versions of the reflected entropy have also been proposed, motivated by AdS/CFT [53–55]. It would be interesting to understand each of these quantities in the context of holography, or even toy models such as tensor networks. If the program of understanding quantum gravity by understanding quantum information is to progress, it is crucial that we obtain a more refined understanding of multipartite entanglement measures.

Applications for reflected entropy continuity. Our new bound in Theorem 2.1 has many interesting applications. For example, it might be useful in proving inequalities about S_R that were conjectured in [10]. Indeed, those inequalities might be easier to prove for e.g. the fixed-area states defined in [47, 48]. Holographic CFT states are generally close in trace distance to one fixed-area state. So, bounds on the reflected entropy of one translate to bounds on the reflected entropy of the other. It would be interesting to find other uses for this theorem.

GHZ isn't enough. While we have demonstrated that tripartite entanglement is necessary for the S_R and E_P conjectures, we have not emphasized what type of tripartite entanglement is required. In fact, GHZ entanglement — even a lot of it — does not help. One can show that GHZ entanglement also satisfies $S_R(A:B) = I(A:B)$. (Note that this problem is also not resolved by adding superselection sectors, similar to the α blocks in operator-algebra quantum error correction [47, 48, 56, 57]. These results strongly suggest that the "stabilizerness" of holographic states is very low, which will be discussed in upcoming work [58].[7])

Beyond this, there is little we can say. It is difficult to pinpoint what type of entanglement must be present, because there are many inequivalent forms of tripartite entanglement, and the classification is not well understood in general. In the case of three qubits,

[7] We thank Brian Swingle for discussions related to this.

JHEP04(2020)208

there are just two inequivalent forms of entanglement: GHZ and W [59]. For A, B two of the three qubits in a W-state, $S_R(A:B) = 1.49 \log 2$ while $I(A:B) = 0.92 \log 2$, and therefore W-entanglement might be used to alleviate the gap between the MBC (and RSTNs) and holography. Similarly, numerical analyses suggest that $E_P(A:B) \neq \frac{1}{2} I(A:B)$ for such states [26, 41]. It would be interesting understand better the particular kind of tripartite entanglement that is crucial for holography.

Acknowledgments

We thank Ning Bao, Raphael Bousso, Ven Chandrasekaran, Netta Engelhardt, Tom Faulkner, Matthew Headrick, Arvin Shahbazi Moghaddam, Yasunori Nomura and Brian Swingle for helpful discussions and comments. C.A. is supported by the US Department of Energy grants DE-SC0018944 and DE-SC0019127, and also the Simons foundation as a member of the It from Qubit collaboration. This work was supported in part by the Berkeley Center for Theoretical Physics; by the Department of Energy, Office of Science, Office of High Energy Physics under QuantISED Award DE-SC0019380 and under contract DE-AC02-05CH11231; and by the National Science Foundation under grant PHY1820912.

References

[1] M. Van Raamsdonk, *Building up spacetime with quantum entanglement*, *Gen. Rel. Grav.* **42** (2010) 2323 [arXiv:1005.3035] [INSPIRE].

[2] J. Maldacena and L. Susskind, *Cool horizons for entangled black holes*, *Fortsch. Phys.* **61** (2013) 781 [arXiv:1306.0533] [INSPIRE].

[3] S. Ryu and T. Takayanagi, *Aspects of Holographic Entanglement Entropy*, *JHEP* **08** (2006) 045 [hep-th/0605073] [INSPIRE].

[4] S. Ryu and T. Takayanagi, *Holographic derivation of entanglement entropy from AdS/CFT*, *Phys. Rev. Lett.* **96** (2006) 181602 [hep-th/0603001] [INSPIRE].

[5] N. Bao, S. Nezami, H. Ooguri, B. Stoica, J. Sully and M. Walter, *The Holographic Entropy Cone*, *JHEP* **09** (2015) 130 [arXiv:1505.07839] [INSPIRE].

[6] V. Balasubramanian, P. Hayden, A. Maloney, D. Marolf and S.F. Ross, *Multiboundary Wormholes and Holographic Entanglement*, *Class. Quant. Grav.* **31** (2014) 185015 [arXiv:1406.2663] [INSPIRE].

[7] S.X. Cui, P. Hayden, T. He, M. Headrick, B. Stoica and M. Walter, *Bit Threads and Holographic Monogamy*, arXiv:1808.05234 [INSPIRE].

[8] P. Hayden, S. Nezami, X.-L. Qi, N. Thomas, M. Walter and Z. Yang, *Holographic duality from random tensor networks*, *JHEP* **11** (2016) 009 [arXiv:1601.01694] [INSPIRE].

[9] S. Nezami and M. Walter, *Multipartite Entanglement in Stabilizer Tensor Networks*, arXiv:1608.02595 [INSPIRE].

JHEP04(2020)208

[10] S. Dutta and T. Faulkner, *A canonical purification for the entanglement wedge cross-section*, arXiv:1905.00577 [INSPIRE].

[11] T. Takayanagi and K. Umemoto, *Entanglement of purification through holographic duality*, *Nature Phys.* **14** (2018) 573 [arXiv:1708.09393] [INSPIRE].

[12] P. Nguyen, T. Devakul, M.G. Halbasch, M.P. Zaletel and B. Swingle, *Entanglement of purification: from spin chains to holography*, *JHEP* **01** (2018) 098 [arXiv:1709.07424] [INSPIRE].

[13] A. Bhattacharyya, T. Takayanagi and K. Umemoto, *Entanglement of Purification in Free Scalar Field Theories*, *JHEP* **04** (2018) 132 [arXiv:1802.09545] [INSPIRE].

[14] P. Caputa, M. Miyaji, T. Takayanagi and K. Umemoto, *Holographic Entanglement of Purification from Conformal Field Theories*, *Phys. Rev. Lett.* **122** (2019) 111601 [arXiv:1812.05268] [INSPIRE].

[15] N. Bao, A. Chatwin-Davies, J. Pollack and G.N. Remmen, *Towards a Bit Threads Derivation of Holographic Entanglement of Purification*, *JHEP* **07** (2019) 152 [arXiv:1905.04317] [INSPIRE].

[16] F. Pastawski, B. Yoshida, D. Harlow and J. Preskill, *Holographic quantum error-correcting codes: Toy models for the bulk/boundary correspondence*, *JHEP* **06** (2015) 149 [arXiv:1503.06237] [INSPIRE].

[17] N. Engelhardt and A.C. Wall, *Coarse Graining Holographic Black Holes*, *JHEP* **05** (2019) 160 [arXiv:1806.01281] [INSPIRE].

[18] Y. Kusuki and K. Tamaoka, *Entanglement Wedge Cross Section from CFT: Dynamics of Local Operator Quench*, *JHEP* **02** (2020) 017 [arXiv:1909.06790] [INSPIRE].

[19] Y. Kusuki and K. Tamaoka, *Dynamics of Entanglement Wedge Cross Section from Conformal Field Theories*, arXiv:1907.06646 [INSPIRE].

[20] K. Umemoto, *Quantum and Classical Correlations Inside the Entanglement Wedge*, *Phys. Rev.* **D 100** (2019) 126021 [arXiv:1907.12555] [INSPIRE].

[21] M. Fannes, *A continuity property of the entropy density for spin lattice systems*, *Commun. Math. Phys.* **31** (1973) 291.

[22] K.M.R. Audenaert, *Comparisons between quantum state distinguishability measures*, arXiv:1207.1197.

[23] K.M. Audenaert, *A sharp fannes-type inequality for the von neumann entropy*, quant-ph/0610146.

[24] N. Engelhardt and A.C. Wall, *Decoding the Apparent Horizon: Coarse-Grained Holographic Entropy*, *Phys. Rev. Lett.* **121** (2018) 211301 [arXiv:1706.02038] [INSPIRE].

[25] C. Akers, T. Faulkner, S. Lin and P. Rath, to appear.

[26] B.M. Terhal, M. Horodecki, D.W. Leung and D.P. DiVincenzo, *The entanglement of purification*, *J. Math. Phys.* **43** (2002) 4286.

[27] M. Miyaji and T. Takayanagi, *Surface/State Correspondence as a Generalized Holography*, *PTEP* **2015** (2015) 073B03 [arXiv:1503.03542] [INSPIRE].

[28] Y. Nomura, P. Rath and N. Salzetta, *Pulling the Boundary into the Bulk*, *Phys. Rev.* **D 98** (2018) 026010 [arXiv:1805.00523] [INSPIRE].

JHEP04(2020)208

[29] B. Chen, L. Chen and C.-Y. Zhang, *Surface/State correspondence and $T\overline{T}$ deformation*, arXiv:1907.12110 [INSPIRE].

[30] A. Prudenziati, *Geometry of Entanglement*, arXiv:1907.06238 [INSPIRE].

[31] N. Cheng, *Optimized Correlation Measures in Holography*, *Phys. Rev.* **D 101** (2020) 066009 [arXiv:1909.09334] [INSPIRE].

[32] W.-Z. Guo, *Entanglement of purification and disentanglement in CFTs*, *JHEP* **09** (2019) 080 [arXiv:1904.12124] [INSPIRE].

[33] K. Babaei Velni, M.R. Mohammadi Mozaffar and M.H. Vahidinia, *Some Aspects of Entanglement Wedge Cross-Section*, *JHEP* **05** (2019) 200 [arXiv:1903.08490] [INSPIRE].

[34] N. Bao, A. Chatwin-Davies and G.N. Remmen, *Entanglement of Purification and Multiboundary Wormhole Geometries*, *JHEP* **02** (2019) 110 [arXiv:1811.01983] [INSPIRE].

[35] N. Bao and I.F. Halpern, *Conditional and Multipartite Entanglements of Purification and Holography*, *Phys. Rev.* **D 99** (2019) 046010 [arXiv:1805.00476] [INSPIRE].

[36] K. Umemoto and Y. Zhou, *Entanglement of Purification for Multipartite States and its Holographic Dual*, *JHEP* **10** (2018) 152 [arXiv:1805.02625] [INSPIRE].

[37] A. Bhattacharyya, A. Jahn, T. Takayanagi and K. Umemoto, *Entanglement of Purification in Many Body Systems and Symmetry Breaking*, *Phys. Rev. Lett.* **122** (2019) 201601 [arXiv:1902.02369] [INSPIRE].

[38] J. Harper and M. Headrick, *Bit threads and holographic entanglement of purification*, *JHEP* **08** (2019) 101 [arXiv:1906.05970] [INSPIRE].

[39] H. Hirai, K. Tamaoka and T. Yokoya, *Towards Entanglement of Purification for Conformal Field Theories*, *PTEP* **2018** (2018) 063B03 [arXiv:1803.10539] [INSPIRE].

[40] S. Bagchi and A.K. Pati, *Monogamy, polygamy, and other properties of entanglement of purification*, *Phys. Rev.* **A 91** (2015) 042323.

[41] J. Chen and A. Winter, *Non-additivity of the entanglement of purification (beyond reasonable doubt)*, arXiv:1206.1307.

[42] D. Marolf, H. Maxfield, A. Peach and S.F. Ross, *Hot multiboundary wormholes from bipartite entanglement*, *Class. Quant. Grav.* **32** (2015) 215006 [arXiv:1506.04128] [INSPIRE].

[43] A. Peach and S.F. Ross, *Tensor Network Models of Multiboundary Wormholes*, *Class. Quant. Grav.* **34** (2017) 105011 [arXiv:1702.05984] [INSPIRE].

[44] N. Bao, G. Penington, J. Sorce and A.C. Wall, *Holographic Tensor Networks in Full AdS/CFT*, arXiv:1902.10157 [INSPIRE].

[45] N. Bao, G. Penington, J. Sorce and A.C. Wall, *Beyond Toy Models: Distilling Tensor Networks in Full AdS/CFT*, *JHEP* **11** (2019) 069 [arXiv:1812.01171] [INSPIRE].

[46] X. Dong, *The Gravity Dual of Renyi Entropy*, *Nature Commun.* **7** (2016) 12472 [arXiv:1601.06788] [INSPIRE].

[47] C. Akers and P. Rath, *Holographic Renyi Entropy from Quantum Error Correction*, *JHEP* **05** (2019) 052 [arXiv:1811.05171] [INSPIRE].

[48] X. Dong, D. Harlow and D. Marolf, *Flat entanglement spectra in fixed-area states of quantum gravity*, *JHEP* **10** (2019) 240 [arXiv:1811.05382] [INSPIRE].

JHEP04(2020)208

[49] N. Bao, M. Moosa and I. Shehzad, *The holographic dual of Rényi relative entropy*, *JHEP* **08** (2019) 099 [arXiv:1904.08433] [INSPIRE].

[50] K. Tamaoka, *Entanglement Wedge Cross Section from the Dual Density Matrix*, *Phys. Rev. Lett.* **122** (2019) 141601 [arXiv:1809.09109] [INSPIRE].

[51] Y. Kusuki, J. Kudler-Flam and S. Ryu, *Derivation of Holographic Negativity in AdS_3/CFT_2*, *Phys. Rev. Lett.* **123** (2019) 131603 [arXiv:1907.07824] [INSPIRE].

[52] J. Levin, O. DeWolfe and G. Smith, *Correlation measures and distillable entanglement in AdS/CFT*, *Phys. Rev.* **D 101** (2020) 046015 [arXiv:1909.04727] [INSPIRE].

[53] D. Marolf, *CFT sewing as the dual of AdS cut-and-paste*, *JHEP* **02** (2020) 152 [arXiv:1909.09330] [INSPIRE].

[54] N. Bao and N. Cheng, *Multipartite Reflected Entropy*, *JHEP* **10** (2019) 102 [arXiv:1909.03154] [INSPIRE].

[55] J. Chu, R. Qi and Y. Zhou, *Generalizations of Reflected Entropy and the Holographic Dual*, *JHEP* **03** (2020) 151 [arXiv:1909.10456] [INSPIRE].

[56] D. Harlow, *The Ryu-Takayanagi Formula from Quantum Error Correction*, *Commun. Math. Phys.* **354** (2017) 865 [arXiv:1607.03901] [INSPIRE].

[57] X. Dong and D. Marolf, *One-loop universality of holographic codes*, *JHEP* **03** (2020) 191 [arXiv:1910.06329] [INSPIRE].

[58] C. Cao and B. Swingle, to appear.

[59] W. Dür, G. Vidal and J.I. Cirac, *Three qubits can be entangled in two inequivalent ways*, *Phys. Rev.* **A 62** (2000) 062314 [quant-ph/0005115] [INSPIRE].

JHEP04(2020)208

Title

Entanglement density and gravitational thermodynamics

Journal

Physical Review D - Particles, Fields, Gravitation and Cosmology, 91(10)

ISSN

1550-7998

Authors

Bhattacharya, J
Hubeny, VE
Rangamani, M
et al.

Publication Date

2015-05-21

DOI

10.1103/PhysRevD.91.106009

DCPT-14/77, YITP-104, IPMU14-0359

Entanglement density and gravitational thermodynamics

Jyotirmoy Bhattacharya,[1,*] Veronika E. Hubeny,[1,†] Mukund Rangamani,[1,‡] and Tadashi Takayanagi[2,3,§]

[1] *Centre for Particle Theory & Department of Mathematical Sciences, Durham University, South Road, Durham DH1 3LE, United Kingdom*
[2] *Yukawa Institute for Theoretical Physics (YITP), Kyoto University, Kyoto 606-8502, Japan*
[3] *Kavli Institute for the Physics and Mathematics of the Universe (WPI), University of Tokyo, Kashiwa, Chiba 277-8582, Japan*

In an attempt to find a quasi-local measure of quantum entanglement, we introduce the concept of entanglement density in relativistic quantum theories. This density is defined in terms of infinitesimal variations of the region whose entanglement we monitor, and in certain cases can be mapped to the variations of the generating points of the associated domain of dependence. We argue that strong sub-additivity constrains the entanglement density to be positive semi-definite. Examining this density in the holographic context, we map its positivity to a statement of integrated null energy condition in the gravity dual. We further speculate that this may be mapped to a statement analogous to the second law of black hole thermodynamics, for the extremal surface.

I. INTRODUCTION

The holographic AdS/CFT correspondence indicates that the fundamental constituents of spacetime geometry are quanta of a conventional non-gravitational field theory. The precise manner in which these non-gravitational quanta conspire to construct a smooth semiclassical spacetime, however, still remains obscure. Holography is motivated by black hole thermodynamics, which suggests that emergence of gravity can be associated with coarse-graining a la classical thermodynamics [1]. We then seek to understand what is being coarse-grained, and how.

A crucial hint is provided by the fact that AdS/CFT geometrizes quantum entanglement: entanglement entropy (EE) in the CFT is given by the area of a certain extremal surface in the bulk [2–4]. Indeed, the fascinating idea of spacetime geometry being the encoder of the entanglement structure of the quantum state [5–7] hints at potentially deep insights into the workings of quantum gravity.

As a first step, we would like to decipher the dynamical equations of gravity from the these statements. In this regard, EE which motivates the connection to geometry, a-priori presents a complication: it is non-local – even in local QFTs, it is defined on a causal domain. The corresponding bulk quantity depends on the bulk geometry along a codimension-2 extremal surface. To make contact with local gravitational physics, it would be convenient to work with a more localizable construct in the dual CFT.[1]

Inspired by this logic, we propose to study a QFT quantity we call entanglement density. This effectively measures two-body quantum entanglement between two infinitesimally small regions. To motivate its construction, consider a quantum field theory on a (rigid) background spacetime $\mathcal{B}$ which is foliated by spacelike Cauchy surfaces Σ. We pick a region $\mathcal{A} \subset \Sigma$ and construct the reduced density matrix $\rho_{\mathcal{A}}$. The entanglement entropy $S_{\mathcal{A}} = -\mathrm{Tr}\,(\rho_{\mathcal{A}} \log \rho_{\mathcal{A}})$ is the von Neumann entropy of this density matrix, and is a functional of $\partial\mathcal{A}$. We propose to retain locality by examining EE for infinitesimal variations of $\partial\mathcal{A}$ (and hence $\mathcal{A}$). Schematically for a configuration ρ_Σ on the Cauchy slice, we define the double variation:[2]

$$\hat{n}\,(\delta_1\mathcal{A}, \delta_2\mathcal{A}) = \delta_1\,\delta_2\,S_{\mathcal{A}} \tag{1}$$

The construction is pictorially illustrated in Fig. 1.

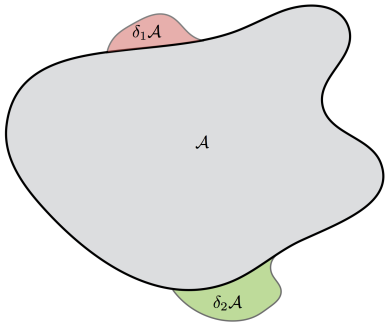


FIG. 1. Illustration of the generic variations $\delta_1\mathcal{A}$ and $\delta_2\mathcal{A}$ which are used to define the entanglement density (1).

Let us now simplify $\hat{n}$ by appealing to the fact that $S_{\mathcal{A}}$ is a functional on the entire domain of dependence $D[\mathcal{A}]$. We focus on backgrounds $\mathcal{B}$ and regions $\mathcal{A}$ for which $D[\mathcal{A}]$ is given by the intersection of past and future light-cones from two points, $\mathcal{C}^{\pm}$ respectively. As a consequence we

* jyotirmoy.bhattacharya@durham.ac.uk
† veronika.hubeny@durham.ac.uk
‡ mukund.rangamani@durham.ac.uk
§ takayana@yukawa.kyoto-u.ac.jp

[1] For recent progress on directly deriving gravitational dynamics from EE, cf., [8–11].

[2] This construction has some parallels with recent discussions of differential entropy introduced in [12] and explored more thoroughly [13].

will focus on the variations inherent in (1) which are due to the variations of one of the points, say $\mathcal{C}^-$, keeping the other fixed (or vice versa).[3] In this context $\delta_1\,\delta_2\,S^{vac}_{\mathcal{A}}=0$ for 2d and 3d CFTs, although (1) pertains in any QFT.

We will exploit the fact that $\hat{n}$ is naturally sensitive to a key property of the von Neumann entropy, namely strong subadditivity (SSA), which states that

$$S_{\mathcal{A}_1\cup\mathcal{A}_2}+S_{\mathcal{A}_1\cap\mathcal{A}_2}\leq S_{\mathcal{A}_1}+S_{\mathcal{A}_2}\qquad\forall\,\mathcal{A}_{1,2}\,. \tag{2}$$

SSA is a convexity property of entanglement; for regions in (2) being small deformations of a parent region, this has a quadratic structure, which motivates (1). Inspired by a beautiful construction of Casini & Huerta [15, 16], we show that entanglement density can be expressed as a second order differential operator $\mathcal{D}^2_{\pm}$ acting on EE by differentiating with respect to the coordinates $\mathsf{c}^{\pm}$ of $\mathcal{C}^{\pm}$ (specified explicitly for $d=2,3$ in §II). SSA then implies $\mathcal{D}^2_{\mathsf{c}^{\pm}}S(\mathsf{c}^+;\mathsf{c}^-)\geq 0$.

Exploiting the holographic construction of EE in terms of bulk codimension-two extremal surfaces $\mathcal{E}_{\mathcal{A}}$, we argue that the variations of interest can be mapped to the motion of the extremal surface along its null normals $N^{\mu}_{(\pm)}$. Using standard differential geometric identities, this in turn can be simplified to a statement about the geometry side of Einstein's equations $E_{\mu\nu}=R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}+\Lambda\,g_{\mu\nu}$, namely

$$\int_{\mathcal{E}_{\mathcal{A}}}\boldsymbol{\epsilon}\;E_{\mu\nu}\,N^{\mu}_{(\pm)}\,N^{\nu}_{(\pm)}\geq 0\,, \tag{3}$$

where $\boldsymbol{\epsilon}$ is the volume form induced on the extremal surface.[4] Indeed, as the main result of this paper we will show that the entanglement density is precisely given by (3) for small perturbations in the AdS_3/CFT_2 setup. We have therefore related SSA (which can be regarded as a physicality condition on EE) to a restriction on the spacetime curvature.[5]

NB: As this work was nearing completion we received [24], where a similar relation between SSA and bulk energy stress tensor has been discussed. Similar results have been obtained by Arias and Casini (unpublished).

II. SSA IN FIELD THEORY

To set the stage for our analysis let us recall the proof of the c-theorem [25] and F-theorem [26–28] based on SSA, as in [15, 16]. We consider subsystems which are defined by the intersection of light-cones from two points $\mathcal{C}^{\pm}$ in d-dimensional QFTs. Letting $D[\mathcal{A}]=J^-[\mathcal{C}^+]\cap J^+[\mathcal{C}^-]$, we pick $\mathcal{A}$ to be a Cauchy slice for $D[\mathcal{A}]$ at constant time; see e.g., Figs 2 and 3. Then $S_{\mathcal{A}}$ can be viewed of as a function of the coordinates $\mathsf{c}^{\pm}$ of $\mathcal{C}^{\pm}$; i.e., $S_{\mathcal{A}}\equiv S(\mathsf{c}^+;\mathsf{c}^-)$. For $\mathcal{B}=\mathbb{R}^{d-1,1}$ we take $\mathsf{c}^{\pm}=(t_{\pm},\mathbf{x}_{\pm})$. Letting $\mathsf{a}=\pm$, we define the entanglement density in $d=2,3$ with respect to varying $\mathcal{C}^{\pm}$ as

$$\hat{n}_{\mathsf{a}}(t_{\mathsf{a}},\mathbf{x}_{\mathsf{a}})\equiv\left[\Box_{\mathsf{a}}+\frac{2\,(d-2)}{t_{\mathsf{a}}}\partial_{t_{\mathsf{a}}}\right]S(t_{\mathsf{a}},\mathbf{x}_{\mathsf{a}})\geq 0\,, \tag{4}$$

where the inequality is guaranteed by SSA. We give a quick overview following [16], with some additional generalizations.

A. QFTs in $d=2$

We start by applying SSA to the configuration in Fig. 2; for space- and time-translation invariant configurations, we can w.l.o.g. fix $\mathcal{C}^+=(0,0)$ as a reference and drop subscripts for coordinates of $\mathcal{C}^-$. SSA implies

$$S_{AD}+S_{CB}\geq S_{AB}+S_{CD}\,. \tag{5}$$

The fact that EE is defined on a causal domain can be used to redefine our region. For example $S_{AD}=S_{AC\,\cup\,CD}$ even for states which are not boost invariant,[6] since both AD and $AC\cup CD$ have the same domain of dependence. As a result we do not make any symmetry assumptions about the state for which EE is evaluated.

Now consider moving $\mathcal{C}^-$ from its original location (t,x) along the light-cone directions to $\mathcal{C}^-_{\nearrow}$ and $\mathcal{C}^-_{\nwarrow}$ respectively by an amount ϵ. This effectively shifts the left and right end-points of $\mathcal{A}$ along the boundary of $D[\mathcal{A}]$ defining the regions on the l.h.s. of (5). For the second region on the r.h.s. we can equivalently consider translating $\mathcal{C}^-\mapsto\mathcal{C}^+_{\uparrow}$ by a distance 2ϵ. Under these shifts we track the implications of SSA (5). In fact, in the present case we simply need to plug in the explicit dependence of the coordinates of the end-points of the various regions:

$$S(t-\epsilon,x-\epsilon)+S(t-\epsilon,x+\epsilon)-S(t,x)-S(t-2\epsilon,x)\geq 0. \tag{6}$$

The inequality (6), upon expanding to second order in ϵ, immediately yields

$$\hat{n}_-\equiv\left(-\partial_t^2+\partial_x^2\right)S(t,x)\geq 0. \tag{7}$$

Repeating the argument with the roles of $\mathcal{C}^{\pm}$ reversed, we obtain $\hat{n}_+\geq 0$.

Note that the inequalities $\hat{n}^{\pm}\geq 0$ can be saturated: as is clear from the relation to the entropic c-function [15], the entanglement densities $\hat{n}_{\pm}$ are vanishing for the

[3] A related version of entanglement density was considered earlier in [8, 14], without invoking the relativistic causal structure.

[4] A sufficient condition for this positivity is the null energy condition. The null energy condition has been crucial in the derivations of SSA [17–19].

[5] For other applications of entropic inequalities and related constraints in gravity duals see [20–23].

[6] Since we have null segments, this statement should be viewed in a suitable limiting sense.

vacuum state of a CFT. Furthermore, they also vanish whenever the EE can be computed in a CFT by a conformal transformation as in [29], which includes, for example, the finite size system at zero temperature and the finite temperature system with an infinitely large size.

Physically, $\hat{n}_\pm$ computes the entanglement between the two infinitesimally small light-like intervals AC and BD in Fig. 2. Since both are directed in the opposite null directions, it is obvious that if the state is completely separated into the left and right-moving sector, the entanglement should be trivial. This explains why the entanglement density is vanishing for ground sates of 2d CFTs. On the other hand, for generic states, for example a ground state of a non-conformal theory, we will find it is non-vanishing.

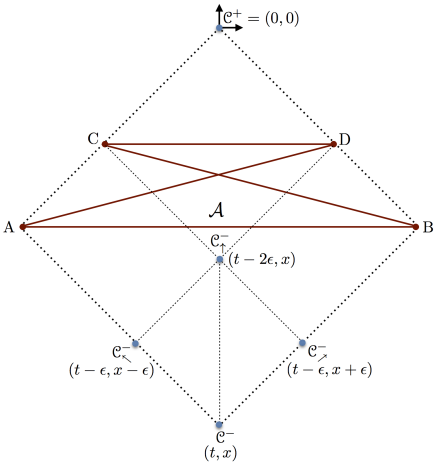


FIG. 2. Illustration of the set-up following [15] in $d = 2$. We choose $\mathcal{C}^+$ to be the origin and the region $\mathcal{A}$ lies on the time-slice with coordinate $\frac{1}{2}t$. We assume $t < 0$ and $\epsilon \leq 0$.

B. QFTs in $d = 3$

The generalization to $d = 3$ can be obtained following [16] by considering the iterated SSA inequality

$$\sum_i S(X_i) \geq S\left(\cup_i X_i\right) + S\left(\cup_{ij}\left(X_i \cap X_j\right)\right) + S\left(\cup_{ijk}\left(X_i \cap X_j \cap X_k\right)\right) + \cdots + S\left(\cap_i X_i\right). \tag{8}$$

We will work in a continuum limit, converting the sums to integrals on both sides of (8).

We once again start with $\mathcal{A}$ defined by $\mathcal{C}^+ = (0, \mathbf{0})$ and $\mathcal{C}^- = (t, \mathbf{x})$. This corresponds to the choice of subsystem given by a round sphere. To apply SSA we consider translating $\mathcal{C}^- \mapsto \mathcal{C}^-_\triangledown$ in the light-cone directions by a distance ϵ, but this time respecting the rotation symmetry. This defines the subsystems X_i, described by ellipses on $\partial D[\mathcal{A}]$. The loci of points composing $\mathcal{C}^-_\triangledown$ is a circle on $\partial J^+[\mathcal{C}^-]$ at time $t - \epsilon$, as indicated in Fig. 3.

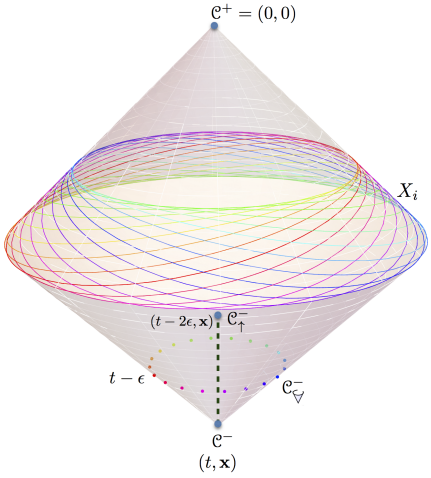


FIG. 3. Illustration of the set-up following [15] in $d \geq 3$ with the same conventions as in Fig. 2. The regions X_i in $d = 3$ are obtained by considering the future light-cone from points distributed on the (dotted) circle, while their iterated intersections are obtained by considering the future light-cone from points on the (dashed) line-segment.

To ascertain the unions of the iterated intersections on the r.h.s. of (8) we make the following observation [16]. Each term in the r.h.s. of (8) generically leads to a curve which averages to a circular cross-section of the light-cone; in the present case we need cross-sections of $\partial J^-[\mathcal{C}^+]$ at constant time. These can equivalently be obtained by translating $\mathcal{C}^- \mapsto \mathcal{C}^-_\uparrow$ in the temporal direction. With this in place we can examine the implications of SSA.

Consider first the contribution from the shift $\mathcal{C}^- \mapsto \mathcal{C}^-_\triangledown$. Writing out the coordinates explicitly we find

$$\text{l.h.s.}_{(8)} = \left[1 - \epsilon\,\partial_t + \frac{\epsilon^2}{4}\left(\nabla^2_{\mathbf{x}} + 2\,\partial_t^2\right)\right] S(t, \mathbf{x}) + \mathcal{O}(\epsilon^3)\,.$$

The r.h.s may be computed similarly, with the only additional complication being that we need to translate the measure from the circular cross-sections of $\partial J^-[\mathcal{C}^+]$ onto the vertical segment along the map $\mathcal{C}^- \mapsto \mathcal{C}^-_\uparrow$. Accounting for this as in [16] we find:

$$\text{r.h.s.}_{(8)} = \left[1 - \epsilon\,\partial_t + \frac{\epsilon^2}{4}\left(3\,\partial_t^2 - \frac{2}{t}\,\partial_t\right)\right] S(t, \mathbf{x}) + \mathcal{O}(\epsilon^3)\,.$$

Combining the above two expressions we have the inequality resulting from SSA:

$$\hat{n}_- \equiv \left[\Box + \frac{2}{t}\partial_t\right] S(t, \mathbf{x}) \geq 0\,. \tag{9}$$

Repeating the analysis about $\mathcal{C}^+$ we can show $\hat{n}_+ \geq 0$. This completes the derivation of (4).

Note that in boost invariant states (e.g., vacuum) where $S_{\mathcal{A}}$ is a function of proper length $\ell = \sqrt{t^2 - ||\mathbf{x}||^2}$,

(4) simply reduces to [15, 16]:

$$\ell\, S''(\ell) - (d-3)\, S'(\ell) \leq 0\,. \tag{10}$$

We have however managed to convert this to a local statement for regions $\mathcal{A}$ which are naturally generated by intersecting light-cones from two points $\mathcal{C}^{\pm}$. Although we have written the expressions (4) and (10) in a manner which suggests an obvious generalization to higher d, there are some subtleties with this interpretation, which we revisit in §IV.

III. HOLOGRAPHIC ENTANGLEMENT DENSITY

Having understood the basic constraint on the entanglement density, let us now consider the holographic context, employing the $\text{AdS}_3/\text{CFT}_2$ duality. We focus on linear perturbations around the pure AdS_3 solution, corresponding to small excitations around the vacuum. In the bulk gravity theory, we consider Einstein gravity coupled to arbitrary matter fields, with the energy-momentum tensor $T_{\mu\nu}$ given by the Einstein's equation

$$E_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda\, g_{\mu\nu} = 8\pi G_N\, T_{\mu\nu}. \tag{11}$$

It is now convenient to work directly with the end-points of $\mathcal{A}$, whose null coordinates are $(u_{_L}, v_{_L})$ and $(u_{_R}, v_{_R})$ respectively in $\mathbb{R}^{1,1}$. In terms of these, the two entanglement densities are given by:

$$\hat{n}_+ = -\frac{\partial}{\partial u_{_R}} \frac{\partial}{\partial v_{_L}} \Delta S_{\mathcal{A}}, \quad \hat{n}_- = -\frac{\partial}{\partial u_{_L}} \frac{\partial}{\partial v_{_R}} \Delta S_{\mathcal{A}}\,. \tag{12}$$

Note that we define the density in terms of $\Delta S_{\mathcal{A}} = S_{\mathcal{A}}^{\rho_\Sigma} - S_{\mathcal{A}}^{vac}$ which measures the entanglement of the excited state ρ_Σ relative to the vacuum. It is crucial here that $\hat{n}_\pm$ vanishes in the vacuum state, for while SSA holds for any state of the CFT, it is no longer true that $\Delta S_{\mathcal{A}}$ satisfies SSA.[7] With this understanding we can replace $S_{\mathcal{A}} \to \Delta S_{\mathcal{A}}$ and still maintain the sign-definiteness of entanglement densities $\hat{n}_\pm$ defined in (12).

We now evaluate $\Delta S_{\mathcal{A}}$ by analyzing the holographic entanglement entropy in the perturbed geometry around pure AdS_3 described by the (gauge fixed) metric:

$$ds^2 = \frac{dz^2 - du\, dv}{z^2} + h_{ab}(u,v,z)\, dx^a\, dx^b\,, \tag{13}$$

where h_{ab} captures the perturbation (Latin indices refer to the boundary). For linear order changes of holographic entanglement entropy, we can work with the original geodesic in AdS_3 (parameterized by ξ) which connects the endpoints of $\mathcal{A}$:

$$(u,v,z) = (U + u_{_\delta} \sin\xi,\ V + v_{_\delta} \sin\xi,\ \sqrt{|u_{_\delta} v_{_\delta}|}\cos\xi)\,,$$

where $\{U, u_{_\delta}\} = \frac{1}{2}(u_{_R} \pm u_{_L})$ and $\{V, v_{_\delta}\} = \frac{1}{2}(v_{_R} \pm v_{_L})$ give the mid-point and separation between the end-points of $\mathcal{A}$.

The first-order perturbation of $\Delta S_{\mathcal{A}}$ is given by

$$\Delta S_A = \frac{1}{8\, G_N} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\xi \frac{\gamma^{(1)}(\xi)}{\sqrt{\gamma^{(0)}(\xi)}}\,, \tag{14}$$

where $\gamma^{(0)}$ and $\gamma^{(1)}$ are induced metric ($\gamma_{\xi\xi}$) at leading and first sub-leading orders, i.e.,

$$\gamma^{(0)}(\xi) = \frac{1}{\cos^2\xi}\,,$$
$$\gamma^{(1)}(\xi) = \cos^2\xi \left(h_{uu}\, u_{_\delta}^2 + h_{vv}\, v_{_\delta}^2 + 2\, h_{uv}\, u_{_\delta}\, v_{_\delta}\right)\,.$$

After some algebra we arrive at the following simple relations:

$$\begin{aligned}\hat{n}_\pm &= \frac{1}{4\, G_N\, |u_{_\delta} v_{_\delta}|} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\xi \sqrt{\gamma^{(0)}} \left(N^\mu_{(\pm)} N^\nu_{(\pm)} E_{\mu\nu}\right) \\ &= \frac{2\pi}{|u_{_\delta} v_{_\delta}|} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\xi \sqrt{\gamma^{(0)}} \left(N^\mu_{(\pm)} N^\nu_{(\pm)} T_{\mu\nu}\right) \geq 0\,,\end{aligned} \tag{15}$$

where $E_{\mu\nu}$ is the l.h.s. of the Einstein's equation (11). The vectors $N^\mu_{(\pm)}$ are the two independent null normals to the extremal surface $\mathcal{E}_{\mathcal{A}}$ in AdS_3,

$$N^\mu_{(\pm)} = \left\{ \frac{u_{_\delta}\, \cos^3\xi}{(\sin\xi \mp 1)}, \frac{v_{_\delta}\, \cos^3\xi}{(\sin\xi \pm 1)}, -\sqrt{|u_{_\delta} v_{_\delta}|}\cos^2\xi \right\}.$$

Firstly, we note from (15) that the positivity of entanglement density is correlated with null energy condition. While we have established the above result explicitly only for linear deviations away from the vacuum, the fact that $\hat{n}_\pm$ vanishes in vacuum, and its positive semi-definiteness from SSA for any excited state, makes it natural for us to conjecture that the relation

$$\hat{n}_\pm = \frac{1}{8\, G_N} \int_{\mathcal{E}_{\mathcal{A}}} d\xi \sqrt{\gamma_{\xi\xi}} \left(N^\mu_{(\pm)} N^\nu_{(\pm)}\, E_{\mu\nu}\right) \geq 0 \tag{16}$$

holds for any asymptotically AdS_3 backgrounds, with $\mathcal{E}_{\mathcal{A}}$ being the extremal surface (spacelike geodesic parameterized ξ) which holographically encodes $S_{\mathcal{A}}$. We leave a more complete exploration of this relation for the future.

It is interesting to note that for normalizable states of pure gravity in AdS_3, the entanglement density always vanishes. This is consistent with our earlier observation that entanglement density is vanishing for any state obtained by conformal transformations of ground states in 2d CFTs. Indeed, solutions in the pure AdS_3 gravity can be obtained by bulk diffeomorphisms corresponding to boundary conformal transformations [30].

IV. DISCUSSION

In this paper we have introduced a new quantity, the entanglement density $\hat{n}$ for relativistic field theories, and

[7] It is easy to verify this statement explicitly say by considering ρ_Σ to be the thermal state.

argued that it provides a useful encoding of certain aspects of gravitational dynamics via holography. We have directly argued for its positivity using the SSA property of EE in 2d and 3d field theories. More generally, we see from our explicit analysis that the positivity of $\hat{n}$ and the gravitational null energy condition go hand in hand. At the same time, we anticipate (16) to be of fundamental importance, since it geometrically encodes the SSA and captures second order variations of holographic entanglement entropy.

While our holographic analysis was carried out for linearized fluctuations around AdS_3, we anticipate that (16) holds at the non-linear level. In fact, it is tempting to conjecture a more general statement valid in any dimension: SSA implies that the entanglement density $\hat{n} \geq 0$ for any state of a QFT with $\hat{n}^{vac} = 0$. Furthermore, translating the description of $\hat{n}$ into holography one finds that (3) holds for any deformation away from pure AdS in arbitrary spacetime dimensions. To wit,

$$\begin{aligned} \text{SSA} \implies & \hat{n}_{\pm} \geq 0\,, \quad \hat{n}^{vac}_{\pm} = 0\,, \\ \implies & \int_{\mathcal{E}_{\mathcal{A}}} \boldsymbol{\epsilon}\, N^{\mu}_{(\pm)} N^{\nu}_{(\pm)}\, E_{\mu\nu} \geq 0\,, \end{aligned} \tag{17}$$

One could try to follow the logic of §II B to arrive at the conclusions above, by considering variations of the past tip of $D[\mathcal{A}]$ (cf., Fig. 3 with each point replaced by $\mathbf{S}^{d-3}$). However, this attempt runs afoul of sub-leading divergences in the entanglement entropy from the r.h.s. of (8) as explained in [16]. It is nevertheless interesting to contemplate whether the entanglement density can be used to provide further insight into c and F-theorems and generalizations thereof.

Nevertheless we may draw the following analogy based on the conjecture above: the statement of SSA is reminiscent of the second law of thermodynamics since it asserts convexity of entanglement (but under region variation as opposed to time variation). We are arguing that this guarantees positivity of the entanglement density. Via holography, generic deformations about the CFT vacuum (equilibrium) then increase the 'cosmological Einstein tensor' $E_{\mu\nu}$ when suitably averaged over the extremal surface. In essence, this quantity codifies a version of gravitational second law for entanglement density. Indeed, in the 'long-wavelength' (hydrodynamic) regime, one may capture the thermal entropy production via the entanglement density by taking $\mathcal{A}$ to be suitably large.

ACKNOWLEDGMENTS

It is a pleasure to thank Shamik Banerjee, Horacio Casini, Matt Headrick, Juan Maldacena, Emil Martinec, Tatsuma Nishioka, and Masahiro Nozaki for discussions.

VH, MR would like to thank the IAS, Princeton, YITP, Kyoto, U. Amsterdam and Aspen Center for Physics (supported by the National Science Foundation under Grant 1066293) for hospitality during the course of this project.

JB is supported by the STFC Consolidated Grant ST/L000407/1. VH and MR were supported in part by the Ambrose Monell foundation, by the FQXi grant "Measures of Holographic Information" (FQXi-RFP3-1334), by the STFC Consolidated Grants ST/J000426/1 and ST/L000407/1, and by the NSF grant under Grant No. PHY-1066293. MR also acknowledges support from the ERC Consolidator Grant Agreement ERC-2013-CoG-615443: SPiN. TT is supported by JSPS Grant-in-Aid for Scientific Research (B) No.25287058 and by JSPS Grant-in-Aid for Challenging Exploratory Research No.24654057. TT is also supported by World Premier International Research Center Initiative (WPI Initiative) from the Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT).

[1] T. Jacobson, Phys.Rev.Lett. **75**, 1260 (1995), gr-qc/9504004.
[2] S. Ryu and T. Takayanagi, JHEP **0608**, 045 (2006), hep-th/0605073.
[3] S. Ryu and T. Takayanagi, Phys.Rev.Lett. **96**, 181602 (2006), hep-th/0603001.
[4] V. E. Hubeny, M. Rangamani, and T. Takayanagi, JHEP **0707**, 062 (2007), 0705.0016.
[5] B. Swingle, Phys.Rev. **D86**, 065007 (2012), 0905.1317.
[6] M. Van Raamsdonk (2009), 0907.2939.
[7] J. Maldacena and L. Susskind, Fortsch.Phys. **61**, 781 (2013), 1306.0533.
[8] M. Nozaki, T. Numasawa, A. Prudenziati, and T. Takayanagi, Phys.Rev. **D88**, 026012 (2013), 1304.7100.
[9] J. Bhattacharya and T. Takayanagi, JHEP **1310**, 219 (2013), 1308.3792.
[10] N. Lashkari, M. B. McDermott, and M. Van Raamsdonk, JHEP **1404**, 195 (2014), 1308.3716.
[11] T. Faulkner, M. Guica, T. Hartman, R. C. Myers, and M. Van Raamsdonk, JHEP **1403**, 051 (2014), 1312.7856.
[12] V. Balasubramanian, B. D. Chowdhury, B. Czech, J. de Boer, and M. P. Heller, Phys.Rev. **D89**, 086004 (2014), 1310.4204.
[13] M. Headrick, R. C. Myers, and J. Wien, JHEP **1410**, 149 (2014), 1408.4770.
[14] M. Nozaki, T. Numasawa, and T. Takayanagi, JHEP **1305**, 080 (2013), 1302.5703.
[15] H. Casini and M. Huerta, Phys.Lett. **B600**, 142 (2004), hep-th/0405111.
[16] H. Casini and M. Huerta, Phys.Rev. **D85**, 125016 (2012), 1202.5650.
[17] A. Allais and E. Tonni, JHEP **1201**, 102 (2012), 1110.1607.
[18] R. Callan, J.-Y. He, and M. Headrick, JHEP **1206**, 081 (2012), 1204.2309.
[19] A. C. Wall, Class.Quant.Grav. **31**, 225007 (2014), 1211.3494.

[20] D. D. Blanco, H. Casini, L.-Y. Hung, and R. C. Myers, JHEP **1308**, 060 (2013), 1305.3182.
[21] S. Banerjee, A. Bhattacharyya, A. Kaviraj, K. Sen, and A. Sinha, JHEP **1405**, 029 (2014), 1401.5089.
[22] R. Bousso, H. Casini, Z. Fisher, and J. Maldacena (2014), 1406.4545.
[23] J. Lin, M. Marcolli, H. Ooguri, and B. Stoica (2014), 1412.1879.
[24] N. Lashkari, C. Rabideau, P. Sabella-Garnier, and M. Van Raamsdonk (2014), 1412.3514.
[25] A. Zamolodchikov, JETP Lett. **43**, 730 (1986).
[26] R. C. Myers and A. Sinha, Phys.Rev. **D82**, 046006 (2010), 1006.1263.
[27] R. C. Myers and A. Sinha, JHEP **1101**, 125 (2011), 1011.5819.
[28] D. L. Jafferis, I. R. Klebanov, S. S. Pufu, and B. R. Safdi, JHEP **1106**, 102 (2011), 1103.1181.
[29] P. Calabrese and J. L. Cardy, J.Stat.Mech. **0406**, P06002 (2004), hep-th/0405152.
[30] K. Skenderis and S. N. Solodukhin, Phys.Lett. **B472**, 316 (2000), hep-th/9910023.

A Study of Entanglement in a Categorical Framework of Natural Language

Dimitri Kartsaklis
University of Oxford
Department of Computer Science
Oxford, UK
dimitri.kartsaklis@cs.ox.ac.uk

Mehrnoosh Sadrzadeh
Queen Mary University of London
School of Electronic Engineering and Computer Science
London, UK
mehrnoosh.sadrzadeh@qmul.ac.uk

In both quantum mechanics and corpus linguistics based on vector spaces, the notion of entanglement provides a means for the various subsystems to communicate with each other. In this paper we examine a number of implementations of the categorical framework of Coecke et al. [4] for natural language, from an entanglement perspective. Specifically, our goal is to better understand in what way the level of entanglement of the relational tensors (or the lack of it) affects the compositional structures in practical situations. Our findings reveal that a number of proposals for verb construction lead to almost separable tensors, a fact that considerably simplifies the interactions between the words. We examine the ramifications of this fact, and we show that the use of Frobenius algebras mitigates the potential problems to a great extent. Finally, we briefly examine a machine learning method that creates verb tensors exhibiting a sufficient level of entanglement.

1 Introduction

Category theory in general and compact closed categories in particular provide a high level framework to identify and study universal properties of mathematical and physical structures. Abramsky and Coecke [1], for example, use the latter to provide a structural proof for a class of quantum protocols, essentially recasting the vector space semantics of quantum mechanics in a more abstract way. This and similar kinds of abstraction have made compact closed categories applicable to other fields with vector space semantics, for the case of this paper, corpus linguistics. Here, Coecke et al.[4] used them to unify two seemingly orthogonal semantic models of natural language: a syntax-driven compositional approach as expressed by Lambek [15] and distributional models of meaning based on vector spaces. The latter approach is capable of providing a concrete representation of the meaning of a word, by creating a vector with co-occurrence counts of that word in a corpus of text with all other words in the vocabulary. Distributional models of this form have been proved useful in many natural language processing tasks [23, 17, 16], but in general they do not scale up to larger text constituents such as phrases and sentences. On the other hand, the type-logical approaches to language as introduced in [15] are compositional but unable to provide a convincing model of word meaning.

The unification of the two semantics paradigms is based on the fact that both a type logic expressed as a pregroup [15] and finite dimensional vector spaces share a compact closed structure; so in principle there exists a way to express a grammatical derivation as a morphism that defines mathematical manipulations between vector spaces, resulting in a sentence vector. In [4], the solution was based on a Cartesian product between the pregroup category and the category of finite dimensional vector spaces; later this was recasted in a functorial passage from the former to the latter [19, 3, 10]. The general idea behind any of these frameworks is that the grammatical type of each word determines the vector space where the corresponding vector lives. Words with atomic types, such as nouns, are simple vectors living in N. On the other hand, words with relational types, such as adjectives or verbs, live in tensor product spaces of higher order. For instance, an intransitive verb will be an element of an order-2 space such as

B. Coecke, I. Hasuo & P. Panangaden (Eds.):
Quantum Physics and Logic 2014 (QPL 2014).
EPTCS 172, 2014, pp. 249–261, doi:10.4204/EPTCS.172.17

$N \otimes S$, whereas a transitive verb will live in $N \otimes S \otimes N$. These tensors act on their arguments by *tensor contraction*, a generalization of the familiar notion of matrix multiplication to higher order tensors.

Since every relational word is represented by a tensor, naturally *entanglement* becomes an important issue in these models. Informally speaking, elements of tensor spaces which represent meanings of relational words should be entangled to allow for a so called 'flow of information' (a terminology borrowed from categorical quantum mechanics [1]) among the meanings of words in a phrase or sentence. Otherwise, parts of the meaning of these words become isolated from the rest, leading to unwanted consequences. An example would be that all sentences that have the same verb end up to get the same meaning regardless of the rest of the context, and this is obviously not the case in language. Whereas at least intuitively the above argument makes sense, in some of the language tasks we have been experimenting with, non-entangled tensors have provided very good results. For example, in [8] Grefenstette and Sadrzadeh provide results for verbs that are built from the outer product of their context vectors. These results beat the state of the art of that time (obtained by the same authors in a previous paper [7]) by a considerable difference.

The purpose of the current paper is to provide a preliminary study of the entanglement in corpus linguistics and to offer some explanation why phenomena such as the above have been the case: is this a by-product of the task or the corpus or the specific concrete model? We work with a number of concrete instantiations of the framework in sentence similarity tasks and observe their performances experimentally from an entanglement point of view. Specifically, we investigate a number of models based on the weighted relations method of [7], where a verb matrix is computed as the structural mixing of all subject/object pairs with which it appears in the training corpus. We also test a model trained using linear regression [2]. Our findings for the first case have been surprising. It turns out that, contrary to intuition and despite the fact that the construction method should yield entangled matrices, the results are very close to their rank-1 approximations, that is, they are in effect separable. We further investigate the ramifications of this observation and try to explain the good practical predictions. We then experiment with the linear regression model of [2] and show that the level of entanglement is much higher in the verbs of this model. Finally, we look at a number of Frobenius variations of the weighted relation models, such as the ones presented in [13] and a few new constructions exclusive to this paper. The conclusions here are also surprising, but in a positive way. It seems that Frobenius models are able to overcome the unwanted "no-flow" collapses of the separable verbs by generating a partial flow between the verb and either its subject or its object, depending which dimension they are copying.

2 Quantizing the grammar

The purpose of the categorical framework is to map a grammatical derivation to some appropriate manipulation between vector spaces. In this section we will shortly review how this goal is achieved. Our basic type logic is a *pregroup grammar* [15], built on the basis of a pregroup algebra. This is a partially ordered monoid with unit 1, whose each element p has a left adjoint p^l and a right adjoint p^r. This means that they satisfy the following inequalities:

$$p^l \cdot p \leq 1 \quad p \cdot p^r \leq 1 \quad \text{and} \quad 1 \leq p \cdot p^l \quad 1 \leq p^r \cdot p \tag{1}$$

A pregroup grammar is the pregroup freely generated over a set of atomic types, which for this paper will be $\{n, s\}$. Here, type n refers to nouns and noun phrases, and type s to sentences. The atomic types and their adjoints can be combined to create types for *relational words*. The type of an adjective, for example, is $n \cdot n^l$, representing something that inputs a noun (from the right) and outputs another noun. Similarly, the type of a transitive verb $n^r \cdot s \cdot n^l$ reflects the fact that verbs of this kind expect two inputs, one noun at each side. A grammatical reduction then follows from the properties of pregroups

and specifically the inequalities in (1) above. The derivation for the sentence 'Happy kids play games' has the following form:

$$(n \cdot n^l) \cdot n \cdot (n^r \cdot s \cdot n^l) \cdot n = n \cdot (n^l \cdot n) \cdot n^r \cdot s \cdot (n^l \cdot n) \leq n \cdot 1 \cdot n^r \cdot s \cdot 1 = n \cdot n^r \cdot s \leq 1 \cdot s = s$$

We refer to the free pregroup generated by a partially ordered set T as $\mathbf{Preg}_F(T)$. Categorically, this structure conforms to the definition of a non-symmetric *compact closed category*. The inequalities in (1) correspond to the ε and η morphisms of a compact closed category, given as follows:

$$\varepsilon^l : A^l \otimes A \to I \qquad \varepsilon^r : A \otimes A^r \to I \tag{2}$$

$$\eta^l : I \to A \otimes A^l \qquad \eta^r : I \to A^r \otimes A \tag{3}$$

Hence the above grammatical reduction becomes the following morphism:

$$(\varepsilon_n^r \otimes 1_s) \circ (1_n \otimes \varepsilon_n^l \otimes 1_{n^r \cdot s} \otimes \varepsilon_n^l) \tag{4}$$

Category $\mathbf{Preg}_F(T)$ is posetal, which means that there is at most one morphism between two given objects. To make this into a full-blown category we work with the free compact closed category generated over T, as described in [20], which we will denote $\mathbf{C}_F(T)$. Furthermore, let us refer to the category of finite-dimensional vector spaces and linear maps over $\mathbb{R}$ as $\mathbf{FVect}_W$, where W is our basic distributional vector space with an orthonormal basis $\{w_i\}_i$. This category is again compact closed (although a symmetric one, since $W \cong W^*$), with the ε and η maps given as follows:

$$\varepsilon^l = \varepsilon^r : W \otimes W \to \mathbb{R} :: \sum_{ij} c_{ij}(\overrightarrow{w_i} \otimes \overrightarrow{w_j}) \mapsto \sum_{ij} c_{ij} \langle \overrightarrow{w_i} | \overrightarrow{w_j} \rangle \tag{5}$$

$$\eta^l = \eta^r : \mathbb{R} \to W \otimes W :: 1 \mapsto \sum_i \overrightarrow{w_i} \otimes \overrightarrow{w_i} \tag{6}$$

The transition from a pregroup reduction to a morphism between vector spaces is achieved by a *strongly monoidal functor* $\mathscr{F} : \mathbf{C}_F(T) \to \mathbf{FVect}_W$ that preserves the compact structure so that $\mathscr{F}(A^l) = \mathscr{F}(A)^l$ and $\mathscr{F}(A^r) = \mathscr{F}(A)^r$. Further, since $\mathbf{FVect}_W$ is symmetric and W has a fixed basis, we have that $\mathscr{F}(A)^r = \mathscr{F}(A)^l \cong \mathscr{F}(A)$. As motivated in previous work [13], we assume that $\mathscr{F}$ assigns the basic vector space W to both of the atomic types, that is we have:

$$\mathscr{F}(n) = \mathscr{F}(s) = W \tag{7}$$

The partial orders between the atomic types are mapped to linear maps from W to W by functoriality. The adjoints of atomic types are also mapped to W, whereas the complex types are mapped to tensor products of vector spaces:

$$\mathscr{F}(n \cdot n^l) = \mathscr{F}(n^r \cdot s) = W \otimes W \qquad \mathscr{F}(n^r \cdot s \cdot n^l) = W \otimes W \otimes W \tag{8}$$

We are now in position to define the meaning of a sentence $w_1 w_2 \ldots w_n$ with type reduction α as follows:

$$\mathscr{F}(\alpha)(\overrightarrow{w_1} \otimes \overrightarrow{w_2} \otimes \ldots \otimes \overrightarrow{w_n}) \tag{9}$$

For example, the meaning of the sentence 'happy kids play games', which has the grammatical reduction (4), is computed as follows:

$$\mathscr{F}\left[(\varepsilon_n^r \otimes 1_s) \circ (1_n \otimes \varepsilon_n^l \otimes 1_{n^r \cdot s} \otimes \varepsilon_n^l)\right] \left(\overrightarrow{happy} \otimes \overrightarrow{kids} \otimes \overrightarrow{play} \otimes \overrightarrow{games}\right) =$$

$$(\varepsilon_W \otimes 1_W) \circ (1_W \otimes \varepsilon_W \otimes 1_{W \otimes W} \otimes \varepsilon_W) \left(\overrightarrow{happy} \otimes \overrightarrow{kids} \otimes \overrightarrow{play} \otimes \overrightarrow{games}\right)$$

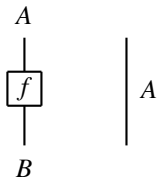
The above categorical computations simplify to the following form:

$$(\overline{happy} \times \overrightarrow{kids})^{\mathsf{T}} \times \overline{play} \times \overrightarrow{games} \tag{10}$$

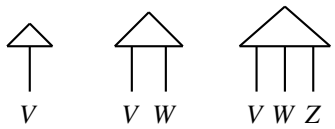
where symbol $\times$ denotes tensor contraction and the above is a vector living in our basic vector space W.

3 Pictorial calculus

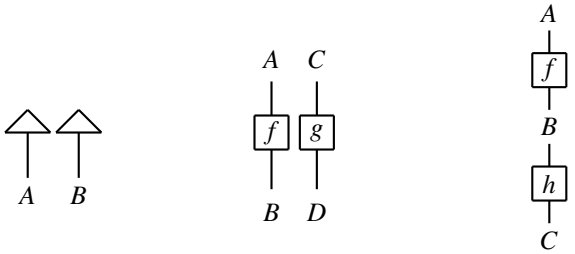
Compact closed categories are complete with regard to a pictorial calculus [14, 24], which can be used for visualizing the derivations and simplifying the computations. We introduce the fragment of calculus that is relevant to the current paper. A morphism $f : A \to B$ is depicted as a box with incoming and outgoing wires representing the objects; the identity morphism $1_A : A \to A$ is a straight line.



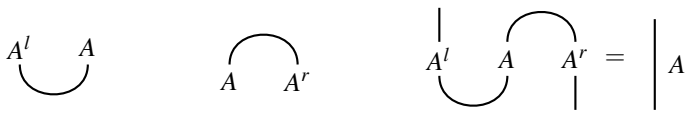
Recall that the objects of $\mathbf{FVect}_W$ are vector spaces. However, for our purposes it is also important to access individual vectors within a vector space. In order to do that, we represent a vector $\overrightarrow{v} \in V$ as a morphism $\overrightarrow{v} : I \to V$. The unit object is depicted as a triangle, while the number of wires emanating from it denotes the order of the corresponding tensor.



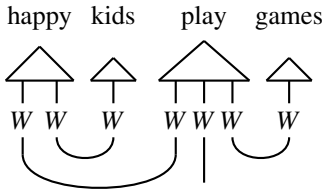
Tensor products of objects and morphisms are depicted by juxtaposing the corresponding diagrams side by side. Composition, on the other hand, is represented as a vertical superposition. For example, from left to right, here are the pictorial representations of the tensor of a vector in A with a vector in B, a tensor of morphisms $f \otimes g : A \otimes C \to B \otimes D$, and a composition of morphisms $h \circ f$ for $f : A \to B$ and $h : B \to C$:



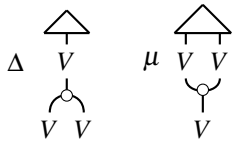
The ε-maps are represented as cups ($\cup$) and the η-maps as caps ($\cap$). Equations such as $(\varepsilon^l_A \otimes 1_{A^r}) \circ (1_{A^l} \otimes \eta^r_A) = 1_A$ now get an intuitive visual justification:



We are now in position to provide a diagram for the meaning of the sentence ‘happy kids play games’.



We conclude this section with one more addition to our calculus. As in most quantum protocols, some times the flow of information in linguistics requires elements of classical processing; specifically, we will want the ability to *copy* and *delete* information, which can be provided by introducing *Frobenius algebras*. In **FVect**, any vector space V with a fixed basis $\{\overrightarrow{v_i}\}$ has a Frobenius algebra over it given by Eqs. 11 below.



$$\Delta :: \overrightarrow{v_i} \mapsto \overrightarrow{v_i} \otimes \overrightarrow{v_i} \qquad \iota :: \overrightarrow{v_i} \mapsto 1 \qquad (11)$$

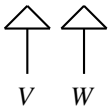
$$\mu :: \overrightarrow{v_i} \otimes \overrightarrow{v_j} \mapsto \delta_{ij} \overrightarrow{v_i} := \begin{cases} \overrightarrow{v_i} & i = j \\ \overrightarrow{0} & i \neq j \end{cases} \qquad \zeta :: 1 \mapsto \sum_i \overrightarrow{v_i}$$

4 Entanglement in quantum mechanics and linguistics

Given two non-interacting quantum systems A and B, where A is in state $|\psi\rangle_A$ and B in state $|\psi\rangle_B$, we denote the state of the composite system $A \otimes B$ by $|\psi\rangle_A \otimes |\psi\rangle_B$. States of this form that can be expressed as the tensor product of two state vectors are called *product* states, and they constitute a special case of separable states. In general, however, the state of a composite system is not necessarily a product state or even a separable one. Fixing bases $\{|i\rangle_A\}$ and $\{|j\rangle_B\}$ for the vector spaces of the two states, a general composite state (separable or not) is denoted as follows:

$$|\psi\rangle_{AB} = \sum_{ij} c_{ij} |i\rangle_A \otimes |j\rangle_B \qquad (12)$$

In the case of a pure quantum state, $|\psi\rangle_{AB}$ is separable only if it can be expressed as the tensor product of two vectors; otherwise it is *entangled*. In a similar way, the tensor of a relational word is separable if it is equal to the tensor product of two vectors. In our graphical calculus, these objects are depicted by the juxtaposition of two or more triangles:

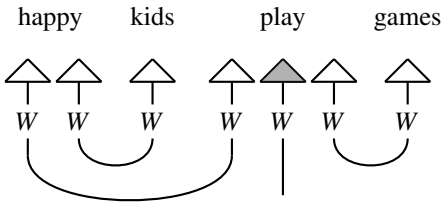


In general, a tensor is not separable if it is a linear combination of many separable tensors. The number of separable tensors needed to express the original tensor is equal to the *tensor rank*. Graphically, a tensor of this form is shown as a single triangle with two or more legs:

5 Consequences of separability

In categorical quantum mechanics terms, entangled states are necessary to allow the flow of information between the different subsystems. In this section we show that the same is true for linguistics. Consider

the diagram of our example derivation, where all relational words are now represented by separable tensors (in other words, no entanglement is present):



In this version, the ε-maps are completely detached from the components of the relational tensors that carry the results (left-hand wire of the adjective and middle wire of the verb); as a consequence, flow of information is obstructed, all compositional interactions have been eliminated, and the meaning of the sentence is reduced to the middle component of the verb (shaded vector) multiplied by a scalar, as follows (superscripts denote the left-hand, middle, and right-hand components of separable tensors):

$$\langle \overrightarrow{happy}^{(r)} | \overrightarrow{kids} \rangle \langle \overrightarrow{happy}^{(l)} | \overrightarrow{play}^{(l)} \rangle \langle \overrightarrow{play}^{(r)} | \overrightarrow{games} \rangle \overrightarrow{play}^{(m)}$$

Depending on how one measures the distance between two sentences, this is a very unwelcome effect, to say the least. When using cosine distance, the meaning of all sentences with 'play' as the verb will be exactly the same and equal to the middle component of the 'play' tensor. For example, the sentence "trembling shadows play hide-and-seek" will have the same meaning as our example sentence. Similarly, the comparison of two arbitrary transitive sentences will be reduced to comparing just the middle components of their verb tensors, completely ignoring any surrounding context. The use of Euclidean distance instead of cosine would slightly improve things, since now we would be at least able to also detect differences in the magnitude between the two middle components. Unfortunately, this metric has been proved not very appropriate for distributional models of meaning, since in the vastness of a highly dimensional space every point ends up to be almost equidistant from all the others. As a result, most implementations of distributional models prefer the more relaxed metric of cosine distance which is length-invariant. Table 1 presents the consequences of separability in a number of grammatical constructs.

Structure	**Simplification**	**Cos-measured result**
adjective-noun	$\overrightarrow{adj} \times \overrightarrow{noun} = (\overrightarrow{adj}^{(l)} \otimes \overrightarrow{adj}^{(r)}) \times \overrightarrow{noun} = \langle \overrightarrow{adj}^{(r)} | \overrightarrow{noun} \rangle \cdot \overrightarrow{adj}^{(l)}$	$\overrightarrow{adj}^{(l)}$
intrans. sentence	$\overrightarrow{subj} \times \overrightarrow{verb} = \overrightarrow{subj} \times (\overrightarrow{verb}^{(l)} \otimes \overrightarrow{verb}^{(r)}) = \langle \overrightarrow{subj} | \overrightarrow{verb}^{(l)} \rangle \cdot \overrightarrow{verb}^{(r)}$	$\overrightarrow{verb}^{(r)}$
verb-object	$\overrightarrow{verb} \times \overrightarrow{obj} = (\overrightarrow{verb}^{(l)} \otimes \overrightarrow{verb}^{(r)}) \times \overrightarrow{obj} = \langle \overrightarrow{verb}^{(r)} | \overrightarrow{obj} \rangle \cdot \overrightarrow{verb}^{(l)}$	$\overrightarrow{verb}^{(l)}$
transitive sentence	$\overrightarrow{subj} \times \overrightarrow{verb} \times \overrightarrow{obj} = \overrightarrow{subj} \times (\overrightarrow{verb}^{(l)} \otimes \overrightarrow{verb}^{(m)} \otimes \overrightarrow{verb}^{(r)}) \times \overrightarrow{obj} = \langle \overrightarrow{subj} | \overrightarrow{verb}^{(l)} \rangle \cdot \langle \overrightarrow{verb}^{(r)} | \overrightarrow{obj} \rangle \cdot \overrightarrow{verb}^{(m)}$	$\overrightarrow{verb}^{(m)}$

Table 1: Consequences of separability in various grammatical structures. Superscripts (l), (m) and (r) refer to left-hand, middle, and right-hand component of a separable tensor

6 Concrete models for verb tensors

Whereas for the vector representations of atomic words of language one can use the much-experimented-with methods of distributional semantics, the tensor representations of relational words is a by-product of the categorical framework whose concrete instantiations are still being investigated. A number of concrete implementations have been proposed so far, e.g. see [7, 13, 9, 12]. These constructions vary from corpus-based methods to machine learning techniques. One problem that researchers have had to address is that tensors of order higher than 2 are difficult to create and manipulate. A transitive verb, for example, is represented by a cuboid living in $W^{\otimes 3}$; if the cardinality of our basic vector space is 1000 (and assuming a standard floating-point representation of 8 bytes per real number), the space required for just a single verb becomes 8 gigabytes. A workaround to this issue is to initially create the verb as a matrix, and then expand it to a tensor of higher order by applying Frobenius Δ operators–that is, leaving one or more dimensions of the resulting tensor empty (filled with zeros).

A simple and intuitive way to create a matrix for a relational word is to structurally mix the arguments with which this word appears in the training corpus [7]. For a transtive verb, this would be given us:

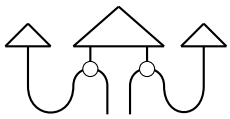
$$\overline{verb} = \sum_i (\overrightarrow{subject_i} \otimes \overrightarrow{object_i}) \tag{13}$$

where $\overrightarrow{subject_i}$ and $\overrightarrow{object_i}$ are the vectors of the subject/object pair for the ith occurrence of the verb in the corpus. The above technique seems to naturally result in an entangled matrix, assuming that the family of subject vectors exhibit a sufficient degree of linear independence, and the same is true for the family of object vectors. Compare this to a straightforward variation which naturally results in a separable matrix, as follows:

$$\overline{verb} = \left(\sum_i \overrightarrow{subject_i} \right) \otimes \left(\sum_i \overrightarrow{object_i} \right) \tag{14}$$

In what follows, we present a number of methods to embed the above verbs from tensors of order 2 to tensors of higher order, as required by the categorical framework.

Relational In [7], the order of a sentence space depends on the arity of the verb of the sentence; for a transitive sentence the result will be a matrix, for an intransitive one it will be a vector, and so on. For the transitive case, the authors expand the original verb matrix to a tensor of order 4 (since now $S = N \otimes N$, the original $N \otimes S \otimes N$ space becomes $N^{\otimes 4}$) by copying both dimensions using Frobenius Δ operators as shown below:

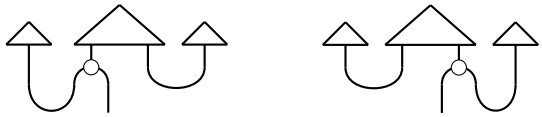


Linear-algebraically, the meaning of a transitive sentence is a matrix living in $W \otimes W$ obtained by the following equation:

$$\overline{subj\ verb\ obj} = (subj \otimes obj) \odot \overline{verb} \tag{15}$$

where the symbol $\odot$ denotes element-wise multiplication.

Frobenius The above method has the limitation that sentences of different structures live in spaces of different tensor orders, so a direct comparison thereof is not possible. As a solution, Kartsaklis et al. [13] propose the copying of only one dimension of the original matrix, which leads to the following two possibilities:



The result is now a vector, computed in the following way, respectively for each case:

$$\textbf{Copy-subject:} \quad \overrightarrow{subj\ verb\ obj} = \overrightarrow{subj} \odot (\overline{verb} \times \overrightarrow{obj}) \tag{16}$$

$$\textbf{Copy-object:} \quad \overrightarrow{subj\ verb\ obj} = \overrightarrow{obj} \odot (\overline{verb}^{\mathsf{T}} \times \overrightarrow{subj}) \tag{17}$$

Each one of the vectors obtained from Eqs. 16 and 17 above addresses a partial interaction of the verb with each argument. It is reasonable then to further combine them in order to get a more complete representation of the verb meaning (and hence the sentence meaning). We therefore define three more models, in which this combination is achieved through vector addition (**Frobenius additive**), element-wise multiplication (**Frobenius multiplicative**), and tensor product (**Frobenius tensored**) of the above.

We conclude this section with two important comments. First, although the use of a matrix for representing a transitive verb might originally seem as a violation of the functorial relation with a pre-group grammar, this is not the case in practice; the functorial relation is restored through the use of the Frobenius operators, which produce a tensor of the correct order, as required by the grammatical type. Furthermore, this notion of "inflation" has the additional advantage that can also work from a reversed perspective: a matrix created by Eq. 13 can be seen as an order-3 tensor originally in $N \otimes S \otimes N$ where the S dimension has been discarded by a ζ Frobenius map. Using this approach, Sadrzadeh and colleagues provide intuitive analyses for wh-movement phenomena and discuss compositional treatments of constructions containing relative pronouns [21, 22].

Finally, we would like to stress out the fact that, despite of the actual level of entanglement in our original verb matrix created by Eq. 13, the use of Frobenius operators as described above equips the inflated verb tensors with an extra level of entanglement in any case. As we will see in Sect. 8 when discussing the results of the experimental work, this detail will be proven very important in practice.

7 Experiments

7.1 Creating a semantic space

Our basic vector space is trained from the ukWaC corpus [5], originally using as a basis the 2,000 content words with the highest frequency (but excluding a list of stop words as well as the 50 most frequent content words since they exhibit low information content). As context we considered a 5-word window from either side of the target word, whereas for our weighting scheme we used local mutual information (i.e. point-wise mutual information multiplied by raw counts). The vector space was normalized and projected onto a 300-dimensional space using singular value decomposition (SVD). These choices are based on our best results in a number of previous experiments [12, 11].

7.2 Detecting sentence similarity

In this section we test the various compositional models of Sect. 6 in two similarity tasks involving pairs of transitive sentences; for each pair, we construct composite vectors for the two sentences, and then we

measure their semantic similarity using cosine distance and Euclidean distance. We then evaluate the correlation of each model's performance with human judgements, using Spearman's ρ. In the first task [7], the sentences to be compared are constructed using the same subject and object and semantically correlated verbs, such as 'spell' and 'write'; for example, 'pupils write letters' is compared with 'pupils spell letters'. The dataset consists of 200 sentence pairs.

We are especially interested in measuring the level of entanglement in our verb matrices as these are created by Eq. 13. In order to achieve that, we compute the *rank-1 approximation* of all verbs in our dataset. Given a verb matrix $\overline{verb}$, we first compute its SVD so that $\overline{verb} = \mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}}$, and then we approximate this matrix by using only the highest eigenvalue and the related left and right singular vectors, so that $\overline{verb}_{R1} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^{\mathsf{T}}$. We compare the composite vectors created by the original matrix (Eq. 13), their rank-1 approximations, and the results of the separable model of Eq. 14. We also use a number of baselines: in the 'verbs-only' model, we compare only the verbs (without composing them with the context), while in the additive and multiplicative models we construct the sentence vectors by simply adding and element-wise multiplying the distributional vectors of their words.

The results (Table 2) revealed a striking similarity in the performances of the entangled and separable versions. Using cosine distance, all three models (relational, rank-1 approximation, separable model) have essentially the same behaviour; with Euclidean distance, the relational model performs again the same as its rank-1 approximation, while this time the separable model is lower.

Model	ρ **with cos**	ρ **with Eucl.**
Verbs only	0.329	0.138
Additive	0.234	0.142
Multiplicative	0.095	0.024
Relational	0.400	0.149
Rank-1 approx. of relational	0.402	0.149
Separable	0.401	0.090
Copy-subject	0.379	0.115
Copy-object	0.381	0.094
Frobenius additive	**0.405**	0.125
Frobenius multiplicative	0.338	0.034
Frobenius tensored	**0.415**	0.010
Human agreement	0.60	

Table 2: Results for the first dataset (same subjects/objects, semantically related verbs)

.

The inevitable conclusion that Eq. 13 actually produces a separable matrix was further confirmed by an additional experiment: we calculated the average cosine similarity of the original matrices with their rank-1 approximations, a computation that revealed a similarity as high as 0.99. Since this result might obviously depend on the form of the noun vectors used for creating the verb matrix, this last experiment was repeated with a number of variations of our basic vector space, getting in every case similarities between verb matrices and their rank-1 approximations higher than 0.97. The observed behaviour can only be explained with the presence of a very high level of linear dependence between the subject vectors and between the object vectors. If every subject vector can be expressed as a linear combination of a small number of other vectors (and the same is true for the family of object vectors), then this would drastically reduce the entanglement of the matrix to the level that it is in effect separable.

Our observations are also confirmed in the second sentence similarity task. Here, we use a variation of one of the datasets in [12], consisting of 108 pairs of transitive sentences. The difference with our first task is that now the sentences of a pair are unrelated in a word level, i.e. subjects, objects, and verbs are all different. The results for this second experiment are presented in Table 3.

Model	ρ with cos	ρ with Eucl.
Verbs only	0.449	0.392
Additive	0.581	0.542
Multiplicative	0.287	0.109
Relational	0.334	0.173
Rank-1 approx. of relational	0.333	0.175
Separable	0.332	0.105
Copy-subject	0.427	0.096
Copy-object	0.198	0.144
Frobenius additive	0.428	0.117
Frobenius multiplicative	0.302	0.041
Frobenius tensored	0.332	0.042
Human agreement	0.66	

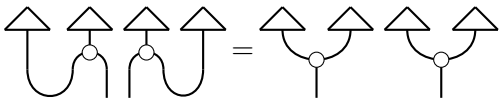
Table 3: Results for the second dataset (different subjects, objects and verbs)

.

As a general observation about the performance of the various models in the two tasks, we note the high scores achieved by the Frobenius models when one uses the preferred method of measurement, that of cosine similarity. Especially the **Frobenius additive** has been proved to perform better than the Relational model, having the additional advantage that it allows comparison between sentences of different structures (since every sentence vector lives in W).

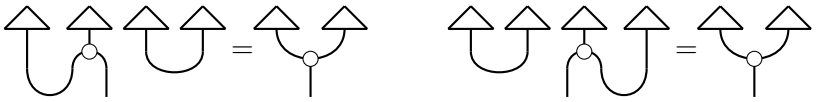
8 Discussion

The experiments of Sect. 7 revealed an unwelcome property of a method our colleagues and we have used in the past for creating verb tensors in the context of compositional models [7, 13, 12]. The fact that the verb matrix is in effect separable introduces a number of simplifications in the models presented in Sect. 6. More specifically, the Relational model of [7] is reduced to the following:



$$\overrightarrow{subj\ verb\ obj} = (\overrightarrow{subj} \odot \overrightarrow{verb}^{(l)}) \otimes (\overrightarrow{verb}^{(r)} \odot \overrightarrow{obj})$$

Furthermore, the Frobenius models of [13] get these forms:



which means, for example, that the actual equation behind the successful Frobenius additive model is

$$\overrightarrow{subj\ verb\ obj} = (\overrightarrow{subj} \odot \overrightarrow{verb}^{(l)}) + (\overrightarrow{verb}^{(r)} \odot \overrightarrow{obj}) \tag{18}$$

Despite the simplifications presented above, note that none of these models degenerates to the level of producing "constant" vectors or matrices, as argued for in Sect. 5. Indeed, especially in the first task (Table 2) the Frobenius models present top performance, and the relational models follow closely. The reason behind this lies in the use of Frobenius Δ operators for copying the original dimensions of the verb matrix, a computation that equipped the fragmented system with flow, although not in the originally intended sense. The compositional structure is still fragmented into two parts, but at least now the copied dimensions provide a means to deliver the results of the two individual computations that take place, one for the left-hand part of the sentence and one for the right-hand part. Let us see what happens when

we use cosine distance in order to compare the matrices of two transitive sentences created with the **Relational** model (the separable version of a verb matrix $\overline{verb}$ is denoted by $\overrightarrow{verb}^{(l)} \otimes \overrightarrow{verb}^{(r)}$):

$$\begin{aligned} \left\langle \overline{subj_1\ verb_1\ obj_1} \middle| \overline{subj_2\ verb_2\ obj_2} \right\rangle &= \\ \left\langle (\overrightarrow{subj}_1 \odot \overrightarrow{verb}_1^{(l)}) \otimes (\overrightarrow{verb}_1^{(r)} \odot \overrightarrow{obj}_1) \middle| (\overrightarrow{subj}_2 \odot \overrightarrow{verb}_2^{(l)}) \otimes (\overrightarrow{verb}_2^{(r)} \odot \overrightarrow{obj}_2) \right\rangle &= \\ \left\langle \overrightarrow{subj}_1 \odot \overrightarrow{verb}_1^{(l)} \middle| \overrightarrow{subj}_2 \odot \overrightarrow{verb}_2^{(l)} \right\rangle \left\langle \overrightarrow{verb}_1^{(r)} \odot \overrightarrow{obj}_1 \middle| \overrightarrow{verb}_2^{(r)} \odot \overrightarrow{obj}_2 \right\rangle & \end{aligned}$$

As also computed and pointed out in [6], the two sentences are broken up to a left-hand part and a right-hand part, and two distinct comparisons take place. As long as we compare sentences of the same structure, as we did here, this method is viable. On the other hand, the **Frobenius** models and their simplifications such as the one in (18) do not have this restriction; in principle, all sentences are represented by vectors living in the same space, so any kind of comparison is possible. In case, however, we do compare sentences of the same structure, these models have the additional advantage that also allow comparisons between *different* sentence parts; this can be seen in the dot product of two sentences created by Eq. 18, which gets the following form:

$$\begin{aligned} &\left\langle \overrightarrow{subj}_1 \odot \overrightarrow{verb}_1^{(l)} \middle| \overrightarrow{subj}_2 \odot \overrightarrow{verb}_2^{(l)} \right\rangle + \left\langle \overrightarrow{subj}_1 \odot \overrightarrow{verb}_1^{(l)} \middle| \overrightarrow{verb}_2^{(r)} \odot \overrightarrow{obj}_2 \right\rangle \quad + \\ &\left\langle \overrightarrow{verb}_1^{(r)} \odot \overrightarrow{obj}_1 \middle| \overrightarrow{subj}_2 \odot \overrightarrow{verb}_2^{(l)} \right\rangle + \left\langle \overrightarrow{verb}_1^{(r)} \odot \overrightarrow{obj}_1 \middle| \overrightarrow{verb}_2^{(r)} \odot \overrightarrow{obj}_2 \right\rangle \end{aligned}$$

9 Using linear regression for entanglement

Corpus-based methods for creating tensors of relational words, such as the models presented so far in this paper, are intuitive and easy to implement. As our experimental work shows, however, this convenience comes with a price. In practice, one would expect that more robust machine learning techniques would produce more reliable tensor representations for composition.

In this section we apply linear regression (following [2]) in order to train verb matrices for a variation of our second experiment, in which we compare elementary verb phrases of the form *verb-object* [18] (so the subjects are dropped). In order to create a matrix for, say, the verb 'play', we first collect all instances of the verb occurring with some object in the training corpus, and then we create non-compositional holistic vectors for these elementary verb phrases following exactly the same methodology as if they were words. We now have a dataset with instances of the form $\langle \overrightarrow{obj}_i, \overrightarrow{play\ obj}_i \rangle$ (e.g. the vector of 'flute' paired with the holistic vector of 'play flute', and so on), that can be used to train a linear regression model in order to produce an appropriate matrix for verb 'play'. The premise of a model like this is that the multiplication of the verb matrix with the vector of a new object will produce a result that approximates the distributional behaviour of all these elementary two-word exemplars used in training. For a given verb, this is achieved by using *gradient descent* in order to minimize the total error between the observed vectors and the vectors predicted by the model, expressed by the following quantity:

$$\frac{1}{2m} \left(\sum_i \overline{verb} \times \overrightarrow{object}_i - \overrightarrow{verb\ object}_i \right)^2 \tag{19}$$

where m is the number of training instances. The average cosine similarity between the matrices we got from this method and their rank-1 approximation was only 0.48, showing that in general the level of

entanglement produced by this method is reasonably high. This is also confirmed by the results in Table 4; the rank-1 approximation model presents the worst performance, since, as you might recall from the discussion in Sect. 5, separability here means that every verb-object composition is reduced to the left component of the verb matrix, completely ignoring the interaction with the object.

Model	ρ **with cos**	ρ **with Eucl.**
Verbs only	0.331	0.267
Holistic verb-phrase vectors	0.403	0.214
Additive	0.379	0.385
Multiplicative	0.301	0.229
Linear regression	0.349	0.144
Rank-1 approximation of LR matrices	0.119	0.082
Human agreement	0.55	

Table 4: Results for the verb-phrase similarity task

10 Conclusion

The current study takes a closer look to an aspect of tensor-based compositional models of meaning that so far had escaped the attention of researchers. The discovery that a number of concrete instantiations of the categorical framework proposed in [4] produce relational tensors that are in effect separable stresses the necessity of similar tests for any linear model that follows the same philosophy. Another contribution of this work was that it showed this is not necessarily a bad thing. The involvement of Frobenius operators in the creation of verb tensors equips the compositional structure with the necessary flow, so that a comparison between two sentence vectors can be still carried out between individual parts of each sentence. Therefore, approaches such as the Frobenius additive model proposed in this paper can be still considered as viable and "easy" alternatives to more robust machine learning techniques, such as the gradient optimization technique discussed in Sect. 9.

References

[1] Samson Abramsky & Bob Coecke (2004): *A Categorical Semantics of Quantum Protocols*. In: *19th Annual IEEE Symposium on Logic in Computer Science*, pp. 415–425, doi:10.1109/LICS.2004.1319636.

[2] M. Baroni & R. Zamparelli (2010): *Nouns are Vectors, Adjectives are Matrices*. In: *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[3] B. Coecke, E. Grefenstette & M. Sadrzadeh (2013): *Lambek vs. Lambek: Functorial Vector Space Semantics and String Diagrams for Lambek Calculus*. *Annals of Pure and Applied Logic*, doi:10.1016/j.apal.2013.05.009. Available at `http://arxiv.org/abs/1302.0393`.

[4] B. Coecke, M. Sadrzadeh & S. Clark (2010): *Mathematical Foundations for Distributed Compositional Model of Meaning. Lambek Festschrift*. *Linguistic Analysis* 36, pp. 345–384.

[5] Adriano Ferraresi, Eros Zanchetta, Marco Baroni & Silvia Bernardini (2008): *Introducing and evaluating ukWaC, a very large web-derived corpus of English*. In: *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pp. 47–54.

[6] E. Grefenstette & M. Sadrzadeh: *Concrete Models and Empirical Evaluations for the Categorical Compositional Distributional Model of Meaning*. *Computational Linguistics*. To appear.

[7] E. Grefenstette & M. Sadrzadeh (2011): *Experimental Support for a Categorical Compositional Distributional Model of Meaning*. In: *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[8] E. Grefenstette & M. Sadrzadeh (2011): *Experimenting with Transitive Verbs in a DisCoCat*. In: *Proceedings of the 2011 EMNLP Workshop on Geometric Models of Natural Language Semantics*.

[9] Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh & Marco Baroni (2013): *Multi-Step Regression Learning for Compositional Distributional Semantics*. In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. Available at `http://arxiv.org/abs/1301.6939`.

[10] D. Kartsaklis, M. Sadrzadeh, S. Pulman & B. Coecke (2014): *Reasoning about Meaning in Natural Language with Compact Closed Categories and Frobenius Algebras*. In J. Chubb, A. Eskandarian & V. Harizanov, editors: *Logic and Algebraic Structures in Quantum Computing and Information*, Association for Symbolic Logic Lecture Notes in Logic, Cambridge University Press. To appear.

[11] Dimitri Kartsaklis, Nal Kalchbrenner & Mehrnoosh Sadrzadeh (2014): *Resolving Lexical Ambiguity in Tensor Regression Models of Meaning*. In: *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers*, Baltimore, USA.

[12] Dimitri Kartsaklis & Mehrnoosh Sadrzadeh (2013): *Prior Disambiguation of Word Tensors for Constructing Sentence Vectors*. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNL)*, Seattle, USA.

[13] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh & Stephen Pulman (2012): *A Unified Sentence Space for Categorical Distributional-Compositional Semantics: Theory and Experiments*. In: *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012): Posters*, The COLING 2012 Organizing Committee, Mumbai, India, pp. 549–558.

[14] G Maxwell Kelly (1972): *Many-Variable Functorial Calculus (I)*. In G.M. Kelly, M. Laplaza, G. Lewis & S. MacLane, editors: *Coherence in categories*, Springer, pp. 66–105, doi:10.1007/BFb0059556.

[15] J. Lambek (2008): *From Word to Sentence*. Polimetrica, Milan.

[16] T. Landauer & S. Dumais (1997): *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquision, Induction, and Representation of Knowledge*. *Psychological Review*, doi:10.1037/0033-295X.104.2.211.

[17] C.D. Manning, P. Raghavan & H. Schütze (2008): *Introduction to Information Retrieval*. Cambridge University Press, doi:10.1017/CBO9780511809071.

[18] Jeff Mitchell & Mirella Lapata (2010): *Composition in Distributional Models of Semantics*. *Cognitive Science* 34(8), pp. 1388–1439, doi:10.1111/j.1551-6709.2010.01106.x.

[19] A. Preller & M. Sadrzadeh (2010): *Bell States and Negative Sentences in the Distributed Model of Meaning*. In P. Selinger B. Coecke, P. Panangaden, editor: *Electronic Notes in Theoretical Computer Science, Proceedings of the 6th QPL Workshop on Quantum Physics and Logic*, University of Oxford, doi:10.1016/j.entcs.2011.01.028.

[20] Anne Preller & Joachim Lambek (2007): *Free compact 2-categories*. *Mathematical Structures in Computer Science* 17(2), pp. 309–340, doi:10.1017/S0960129506005901.

[21] Mehrnoosh Sadrzadeh, Stephen Clark & Bob Coecke (2013): *The Frobenius Anatomy of Word Meanings I: Subject and Object Relative Pronouns*. *Journal of Logic and Computation* 23(6), pp. 1293–1317, doi:10.1093/logcom/ext044.

[22] Mehrnoosh Sadrzadeh, Stephen Clark & Bob Coecke (2014): *The Frobenius Anatomy of Word Meanings II: Possessive Relative Pronouns*. *Journal of Logic and Computation*, doi:10.1093/logcom/exu027.

[23] H. Schütze (1998): *Automatic Word Sense Discrimination*. *Computational Linguistics* 24, pp. 97–123.

[24] Peter Selinger (2011): *A Survey of Graphical Languages for Monoidal Categories*. In Bob Coecke, editor: *New structures for physics*, Springer, pp. 289–355.

Investigating student understanding of quantum entanglement

Antje Kohnle and Erica Deffebach
University of St Andrews, School of Physics and Astronomy,
North Haugh, St Andrews, KY16 9SS, United Kingdom

Abstract. Quantum entanglement is a central concept of quantum theory for multiple particles. Entanglement played an important role in the development of the foundations of the theory and makes possible modern applications in quantum information technology. As part of the QuVis Quantum Mechanics Visualization Project, we developed an interactive simulation *Entanglement: The nature of quantum correlations* using two-particle entangled spin states. We investigated student understanding of entanglement at the introductory and advanced undergraduate levels by collecting student activity and post-test responses using two versions of the simulation and carrying out a small number of student interviews. Common incorrect ideas found include statements that all entangled states must be maximally entangled (i.e. show perfect correlations or anticorrelations along all common measurement axes), that the spins of particles in a product state must have definite values (cannot be in a superposition state with respect to spin) and difficulty factorizing product states. Outcomes from this work will inform further development of the QuVis *Entanglement* simulation.

PACS: 01.50.ht, 03.65.Ud

I. INTRODUCTION

For classical composite systems each of the subsystems has well-defined properties. For quantum-mechanical composite systems, there exist states for which the wave function of the composite system is known, but the subsystems cannot be described in terms of individual wave functions and thus cannot be described separately. Such states for which the total wave function is not the product of individual wave functions, e.g. is not factorizable, are called entangled. Thus, entangled states are not product states.

Schrödinger famously stated that "entanglement is not one but rather the characteristic trait of quantum mechanics" [1]. A remarkable feature is that two entangled quantum particles can show correlations in measurement outcomes that are not reproducible by classical models. Through this feature, entanglement has important physical consequences including the Bell inequalities and applications in teleportation, quantum computing and cryptography [2-4].

Given the key role of entanglement in the description of quantum systems of multiple particles, helping students come to a correct understanding of entanglement is an important instructional goal. Existing studies of student difficulties in quantum mechanics cover various topics but do not include entanglement [5]. As part of the QuVis Quantum Mechanics Visualization Project [6], we have developed an interactive simulation *Entanglement: The nature of quantum correlations* (henceforth referred to as the *Entanglement* simulation) using two-particle entangled spin states. The simulation allows students to explore experimental outcomes for various input states and easily switch between product states and entangled states [7]. In this study, we investigated student understanding of entanglement using two versions of the simulation. Our aims in this work are to assess what common incorrect ideas persist after instruction and *Entanglement* simulation use. Outcomes will inform further development of this simulation.

II. METHODOLOGY

The QuVis *Entanglement* simulation does not require the mathematical formalism of tensor products and is aimed at the introductory and advanced undergraduate levels. A screenshot of the revised, second version of the simulation is shown in Fig. 1. The simulation shows a source of particle pairs in the middle of two Stern-Gerlach apparatuses (SGAs), which can be jointly rotated along two orthogonal axes, denoted X and Z. The states $|\uparrow_A\rangle$ and $|\downarrow_B\rangle$ refer to spin-up and spin-down states along the Z-axis for particles A and B respectively. Students can choose between different input states (left panel in Fig. 1) and send particle pairs through the experiment. The individual and paired measurement outcomes and the correlation coefficient are shown (middle and right panels in Fig. 1). The correlation coefficient is the average value of the product of the two measurement outcomes, defined as +1 when the deflections are the same and –1 when the deflections are opposite. A correlation coefficient of +1 implies perfect correlation, of –1 perfect anticorrelation. Besides the "Controls" view shown in Fig. 1, the simulation also includes explanatory texts in the "Introduction" and "Step-by-step Explanation" views.

In the initial first version of the simulation, users could only choose between three fixed input states, including one

2015 PERC Proceedings, edited by Churukian, Jones, and Ding; Peer-reviewed, doi:10.1119/perc.2015.pr.038

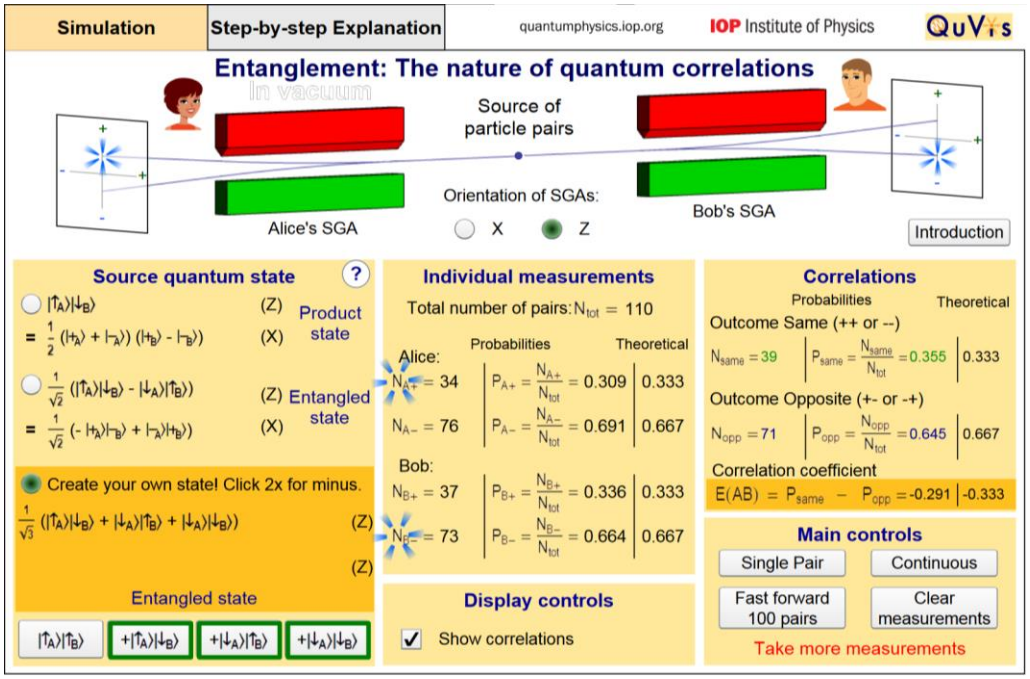


FIG 1. A screenshot of the "Controls" view of the revised version of the *Entanglement* simulation.

entangled state (see Fig. 2). Users could choose to display the states in the X or Z basis, and show the states as products of the two individual particle states or in expanded form as depicted in Fig. 2. Due to difficulties found (see section Outcomes), the revised simulation shows the first two states in both the X and Z bases. It allows users to create their own state by putting together different two-particle spin states, as shown in the lower-left panel of Fig. 1. The revised version also allows users to choose between two different notations for the spin states.

The accompanying activity to the revised simulation shown in Fig. 1 asks students to explain the observed individual and paired measurement outcomes and the correlation coefficient for the first input state (a product state) considering both orientations of the Stern-Gerlach apparatuses. Students are asked to rewrite this state in the X basis and explain why this state is a product state. The activity then asks students to choose the second input state (a maximally entangled state that always has opposite outcomes), and to compare and contrast the previous product state and this entangled state in terms of measurement outcomes. Students are then asked to use the "Create your own state" option to create entangled states with different correlations, including an entangled state for which there are no correlations in the X and Z bases. Students are also asked whether a product state implies that the spins of particles have definite values. The activity to the original version of the simulation was similar, but did not include the parts where students create their own states as this option was not available (see Fig. 2).

We collected written responses to the *Entanglement* simulation activity and in cases also written post-test responses using the original and the revised versions of the simulation (see Table 1). The 2015 post-test questions are shown in Fig. 3. For post-test question 1, states a) and d) are entangled. For post-test question 2, only statement II is correct. The post-test questions are multiple-choice, but students were asked to explain their reasoning for each question. Trials using the simulation were carried out in an introductory quantum physics course (often the first university course in quantum physics that students take, similar to a US Modern Physics course) and a senior-level Advanced Quantum Mechanics course, both at the University of St Andrews.

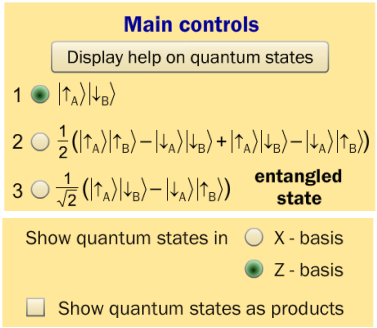


FIG 2. A screenshot showing parts of the original version of the *Entanglement* simulation. Only three fixed input states are available. States are shown in either the X or Z basis, not both simultaneously.

Revisions were incorporated into the simulation prior to the 2015 trial based on analysis of the 2013 and 2014 trials. For the advanced course, post-tests were given in the

TABLE 1. The table shows the number N of activity and post-test responses where applicable collected from courses at the University of St Andrews. The 2013 and 2014 courses used the initial version of the simulation, the 2015 course used the revised version.

Level	Year	N	Post-test	Simulation use
Introductory	2013	59	none	Computer workshop
Advanced	2014	24	Post-test	Homework
Introductory	2015	79	Post-test	Computer workshop

lecture directly after homework submission. For the introductory course, the post-test was completed in the last minutes of a 50-minute computer workshop. We also carried out interviews in 2015 with five students from the introductory level a few days after the *Entanglement* simulation was used. These interviews confirmed our interpretation of written student reasoning.

1) Consider the following two-particle states of a pair of spin ½ particles. The notations $|\uparrow\rangle$ and $|\downarrow\rangle$ refer to spin states with z-component of spin $S_z = +\hbar/2$ and $S_z = -\hbar/2$ respectively. The numerical factors are needed for normalization. Which of these states is/are entangled? Choose one or more.

a) $1/\sqrt{2}\,(|\uparrow_A\rangle|\uparrow_B\rangle + |\downarrow_A\rangle|\downarrow_B\rangle)$

b) $1/2\,(|\uparrow_A\rangle|\uparrow_B\rangle + |\downarrow_A\rangle|\downarrow_B\rangle + |\uparrow_A\rangle|\downarrow_B\rangle + |\downarrow_A\rangle|\uparrow_B\rangle)$

c) $1/\sqrt{2}\,(|\downarrow_A\rangle|\uparrow_B\rangle - |\downarrow_A\rangle|\downarrow_B\rangle)$

d) $1/2\,(|\uparrow_A\rangle|\uparrow_B\rangle + |\downarrow_A\rangle|\downarrow_B\rangle - |\uparrow_A\rangle|\downarrow_B\rangle + |\downarrow_A\rangle|\uparrow_B\rangle)$

2) Do you agree or disagree with the following statements:

I. If the source is emitting entangled particle pairs, then there will be perfect correlation along both X and Z, or perfect anticorrelation along both X and Z.

II. If Alice and Bob find perfect anticorrelation along both X and Z, they know that the source must be emitting entangled particle pairs.

FIG 3. The 2015 post-test questions. Question 1 of the 2014 post-test only included options a), b) and c). Question 2 was only used in 2015.

We marked written responses to the activity questions as correct, partially correct, incorrect and unanswered and compiled the fractions of each per question. For the post-tests, we analyzed students' choices and reasoning in assessing correctness of responses and incorrect ideas, with both reasoning and choices needing to be correct for a response to be coded as correct. We coded incorrect and partially correct responses using an emergent coding scheme, using the same codes for the activity responses and the post-test responses. The 2013 and 2014 activity responses and post-test responses including reasoning were coded by both authors and checked for inter-rater reliability. Categories with disagreement were discussed and revised until high inter-rater reliability was achieved (88% agreement for the 2013 data and 86% for the 2014 data). Due to time constraints, the 2015 data was only coded by one author and checked for consistency by the other author using a subset of the data.

In the lectures, the introductory course only discussed a maximally-entangled two-particle state, and did not define entangled states in terms of not factorizable states. Thus, introductory students were learning about product states and non-maximally entangled states from the simulation alone. In the advanced course lectures, entanglement was introduced via states that are not product states but the focus was primarily on maximally-entangled states and the density matrix formalism.

III. OUTCOMES

In what follows, we discuss common incorrect ideas found in student reasoning. Frequencies across the different levels and years are summarized in Table 2.

A. For an entangled state, if you know the measurement outcome of one particle, the outcome of the other particle is completely determined. Entangled states show either perfect correlations or perfect anticorrelations along all common axes. This idea incorrectly assumes that all entangled states are maximally-entangled, i.e. show either perfect correlation or perfect anticorrelation along all common measurement axes. A typical student response describing entanglement illustrating this idea is "*If you make a measurement on one particle, you know the measurement of the other and they have to be either the opposite or either the same*."

The 2014 post-test question 1 included three states (options a) to c) in Fig. 3, one maximally-entangled state and two product states). Five advanced level students correctly identified the entangled state, but incorrectly reasoned that product states are those for which the outcome of one particle is not fixed when the other is measured. For example, a student reasons "*for a) the two outcomes of the experiment are both A and B measuring the same spin. This means that there is a dependence upon the measurement of A on B and vice versa. Hence a) is an entangled state. For the other states the outcomes can differ in whether A and B measured the same or opposite spin, hence no dependence exists between the measurements. Therefore b) and c) are not entangled*."

These outcomes led us to develop the "Create your own state" option in the revised simulation (see Fig. 1) used in 2015. The revised activity now asks students to create entangled states that do not exhibit perfect correlations or anticorrelations. However, 28 students (35%) in the 2015 introductory level trial incorrectly agreed with post-test question 2 statement I (see Fig. 3), showing that this incorrect idea persists even after students made use of the revised simulation.

B. Incorrect properties of product states, e.g. that product states can be entangled states along a different basis, or that product states can also show perfect

correlations along all bases. These ideas are linked with difficulties translating a state from the Z to the X basis. The 2015 activity explicitly asked students to rewrite a state given in the Z basis in the X basis, and asked "If a state is a product state along Z, will it also be a product state along X?" These two questions were amongst the most poorly answered, with 74% and 78% correct respectively. Also, 23 students (29%) in the 2015 post-test incorrectly disagreed with statement II of question 2 (Fig. 3). Of these students, 11 (14%) used reasoning similar to *"[statement] II is not correct because product states can exist where there is perfect anticorrelation or correlation along X and Z*." The 2013 and 2014 data did not include questions testing for this difficulty.

C. Particles in a product state must have definite spin values (i.e., not be in a superposition of spin states). For the introductory 2013 course, 6 of 59 students (10%) stated that this is the case in response to a question "Entangled states are not product states. Interpret this statement physically." For example, a student states "*For a product state both particles have a definite value of spin measured along a given axis. For an entangled state both particles do not have well-defined spins although their relative spins are always well-defined*." In the 2015 activity to the revised simulation, we explicitly asked "Does a product state imply that the spins of the particles have definite values?" 10 of 79 (13%) students incorrectly stated that this is the case. Several answers stated (not seen in the 2013 responses) that at least one of the particles must have a definite spin. For example, a student states "*It implies that at least one half of the particle pair does.*" This difficulty was only seen at the introductory level.

D. Incorrectly stating that a product state is an entangled state, due to difficulties converting a product state written as a sum of two-particle terms into the factorized form as a product of two single-particle states. In the advanced level course, 5 students (21%) stated on question 1 of the post-test (Fig. 3) that $1/\sqrt{2}\,(|\downarrow_A\rangle|\uparrow_B\rangle - |\downarrow_A\rangle|\downarrow_B\rangle)$ is an entangled state as it could not be factorized, i.e. did not recognize that this is the product state $1/\sqrt{2}\,|\downarrow_A\rangle\,(|\uparrow_B\rangle - |\downarrow_B\rangle)$. In the 2015 post-test 8 students (10%) stated the above state could not be factorized. 13 students (16%) stated that state b) (Fig. 3) is an entangled state as it could not be factorized, whereas this is the product $1/2\,(|\uparrow_A\rangle + |\downarrow_A\rangle)\,(|\uparrow_B\rangle + |\downarrow_B\rangle)$. For the 2015 trial, 27 students in total did not factorize states correctly in the post-test (some responses incorrectly factorized entangled states). The 2013 activity did not include questions assessing this difficulty.

Other difficulties seen with lower frequencies include the incorrect ideas that a quantum state with multiple terms must be an entangled state and that entangled states and mixtures are experimentally indistinguishable. There were also incorrect assignments of correlations to quantum states, e.g. stating that a correlation coefficient of +1 implies the individual outcomes must be completely random.

TABLE 2. Frequencies of common difficulties found; codes as in the text. Student numbers are in parentheses.

Code	Intro 2013 Activity	Advanced 2014 Post-test	Intro 2015 Activity	Intro 2015 Post-test
A	8% (5)	21% (5)	33% (26)	35% (28)
B	0% (0)	0% (0)	16% (13)	14% (11)
C	10% (6)	0% (0)	13% (10)	0% (0)
D	0% (0)	21 % (5)	10% (8)	34% (27)

IV. CONCLUSIONS

These findings point to difficulties with the relations between superposition and entanglement (entanglement implies superposition but not vice versa, code C) and perfect correlations / anticorrelations along multiple axes and entanglement (these correlations imply entanglement but not vice versa, codes A and B). Based on these outcomes, we plan to revise the *Entanglement* simulation to include another view where students can change the coefficients in an entangled state to explore the transition between maximal and non-maximal entanglement. We plan to add help texts showing how to convert between a sum of terms and the factorized form for a product state and to translate a state from the Z to the X basis. We plan to add a "Challenges" view with multiple challenges targeting the difficulties found. Future work will aim to elicit underlying reasons for the difficulties found in this study, and evaluate the effectiveness of these further simulation revisions using pre- and post-tests and student interviews.

ACKNOWLEDGEMENTS

We thank Charles Baily, Christopher Hooley and Natalia Korolkova from the University of St Andrews for incorporating the *Entanglement* simulation into their courses. We gratefully acknowledge all of the students who participated in this study. We thank the UK Institute of Physics for funding the simulation development.

[1] E. Schrödinger, Proc. Cambridge Phil. Soc. **31**, 555 (1935).
[2] J. Bell, Physics **1**, 195 (1964).
[3] C.H. Bennett *et al.*, Phys. Rev. Lett. **70**, 1895 (1993).
[4] C.H. Bennett, G. Brassard and N.D. Mermin, Phys. Rev. Lett. **68**, 557 (1992).
[5] C. Singh and E. Marshman, Phys. Rev. ST Phys. Educ. Res. **11**, 020117 (2015).
[6] www.st-andrews.ac.uk/physics/quvis; A. Kohnle *et al.*, Am. J. Phys. **83**, 560 (2015).
[7] www.st-andrews.ac.uk/physics/quvis/simulations_html5/sims/entanglement/entanglement.html

Vol. 34: 431–448, 2017
https://doi.org/10.3354/esr00865

ENDANGERED SPECIES RESEARCH
Endang Species Res

Published December 11

Contribution to the Special 'Marine pollution and endangered species'

REVIEW

A global review of marine turtle entanglement in anthropogenic debris: a baseline for further action

Emily M. Duncan[1,2,3,*], Zara L. R. Botterell[1,*], Annette C. Broderick[1], Tamara S. Galloway[2], Penelope K. Lindeque[3], Ana Nuno[1], Brendan J. Godley[1,]**

[1]Marine Turtle Research Group, Centre for Ecology and Conservation, University of Exeter, Penryn TR10 9FE, UK
[2]Biosciences, College of Life and Environmental Sciences, Geoffrey Pope Building, University of Exeter, Stocker Road, Exeter EX4 4QD, UK
[3]Marine Ecology and Biodiversity, Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK

ABSTRACT: Entanglement in anthropogenic debris poses a threat to marine wildlife. Although this is recognised as a cause of marine turtle mortality, there remain quantitative knowledge gaps on entanglement rates and population implications. We provide a global summary of this issue in this taxon using a mixed methods approach, including a literature review and expert opinions from conservation scientists and practitioners worldwide. The literature review yielded 23 reports of marine turtle entanglement in anthropogenic debris, which included records for 6 species, in all ocean basins. Our experts reported the occurrence of marine turtles found entangled across all species, life stages and ocean basins, with suggestions of particular vulnerability in pelagic juvenile life stages. Numbers of stranded turtles encountered by our 106 respondents were in the thousands per year, with 5.5% of turtles encountered entangled; 90.6% of these dead. Of our experts questioned, 84% consider that this issue could be causing population level effects in some areas. Lost or discarded fishing materials, known as 'ghost gear', contributed to the majority of reported entanglements with debris from land-based sources in the distinct minority. Surveyed experts rated entanglement a greater threat to marine turtles than oil pollution, climate change and direct exploitation but less of a threat than plastic ingestion and fisheries bycatch. The challenges, research needs and priority actions facing marine turtle entanglement are discussed as pathways to begin to resolve and further understand the issue. Collaboration among stakeholder groups such as strandings networks, the fisheries sector and the scientific community will facilitate the development of mitigating actions.

KEY WORDS: Conservation · Entanglement · Ghost fishing · Marine debris · Plastic pollution · Sea turtle · Strandings

INTRODUCTION

Marine plastic pollution

Anthropogenic materials, the majority of them plastic, are accumulating on the surface of the oceans, in the water column and on the seabed (Thompson et al. 2004). The durability of plastic means that it may persist for centuries (Barnes et al. 2009). It is estimated that 4.8 to 12.7 million tonnes of plastic waste could be entering the marine environment annually (Jambeck et al. 2015). Over 700 marine species have been demonstrated to interact with marine plastic pollution (Gall & Thompson 2015), which presents a risk to ani-

*These authors contributed equally to this work
**Corresponding author: b.j.godley@exeter.ac.uk

Publisher: Inter-Research · www.int-res.com

mals through ingestion, entanglement, degradation of key habitats and wider ecosystem effects (Nelms et al. 2016). Megafauna such as marine turtles with complex life histories and highly mobile behaviour are particularly vulnerable to its impacts (Schuyler et al. 2014).

Entanglement in marine litter

Entanglement in plastic debris is recognised as a major risk for many marine species (Laist 1987, Vegter et al. 2014). This has become sufficiently high profile that the European Union's Marine Strategy Framework Directive (MSFD) Technical Subgroup on Marine Litter has announced that it will develop a dedicated monitoring protocol for its next report (MSFD GES Technical Subgroup on Marine Litter 2011). Entanglement has the potential to cause a range of fatal and non-fatal impacts such as serious wounds leading to maiming, amputation, increased drag, restricted movement or choking (Votier et al. 2011, Barreiros & Raykov 2014, Lawson et al. 2015).

Types of marine debris causing entanglement

The debris causing this entanglement falls into 2 broad categories. Firstly, hundreds of tons of fishing gear are lost, abandoned or discarded annually, forming 'ghost gear' which passively drifts over large distances, sometimes indiscriminately 'fishing' marine organisms (Macfadyen et al. 2009, Wilcox et al. 2013). This gear is commonly made of non-biodegradable synthetic material that will persist in the marine environment, potentially become biofouled by marine organisms and act as a fish aggregating device (FAD), attracting both grazers and predators such as marine turtles (Filmalter et al. 2013, Wilcox et al. 2013). It is important to distinguish here between 'entanglement' and 'bycatch'. Bycatch can be defined as unselective catch of either unused or unmanaged species during fishing, with a particular focus on 'active' gear, whereas ghost gear can be defined as equipment of which the fisher has lost operational control (Smolowitz 1978, Davies et al. 2009). Therefore, here we consider animals caught in passive ghost fishing gear as entangled, not bycaught. Secondly, there have also been reports of entanglement in litter from land-based sources (Chatto 1995, Bentivegna 1995, Santos et al. 2015). In this review we do not include bycaught turtles — only those that have become entangled in passive anthropogenic debris such as ghost gear or land-based debris.

Current knowledge gaps regarding turtle entanglement

Despite turtle entanglement being recognised as one of the major sources of turtle mortality in northern Australia and the Mediterranean, there is a quantitative knowledge gap with respect to the entanglement rates and possible implications in terms of global populations (Casale et al. 2010, Wilcox et al. 2013, Camedda et al. 2014, Gilman et al. 2016). A recent literature review by Nelms et al. (2016) returned only 9 peer-reviewed publications on marine debris entanglement and turtles (Bentivegna 1995, Chatto 1995, López-Jurado et al. 2003, Casale et al. 2010, Santos et al. 2012, Jensen et al. 2013, Wilcox et al. 2013, 2015, Barreiros & Raykov 2014). Of these, 7 were focused on ghost fishing gear, highlighting the distinct lack of knowledge of entanglement in debris from land-based sources. Even fewer of these studies focused on the potential variable susceptibility among life stages or species, with only one paper, Santos et al. (2012), reporting that the majority of entangled olive ridley turtles *Lepidochelys olivacea* on the Brazilian islands of Fernando de Noronha and Atol das Rocas were sub-adults and adults.

Research rationale in terms of marine turtles and pollution

In terms of global research priorities for sea turtle conservation and management, understanding the impact of pollution is considered of high importance (Hamann et al. 2010, Rees et al. 2016). To evaluate this effectively, the impact of anthropogenic debris, specifically, must be considered at a species and population level. Additionally, it is important to understand the variation in entanglement rates among species and life stages to better evaluate vulnerability and the frequency of interactions with different debris types (Nelms et al. 2016). Once these have been established, opportunities for delivering effective education and awareness can be given or other mitigation planned (Vegter et al. 2014).

Here, we define marine turtle entanglement as 'the process under which a marine turtle becomes entwined or trapped within anthropogenic materials.' We sought to include discarded fishing gear (ghost fishing) as well as land-based sources. The aim of this study was to (1) review existing, and obtain new, reports of the occurrence and global spatial distribution of marine turtle entanglement; (2) gain insights into patterns of species, life stage and

debris type involved across entanglement cases; and (3) glean an insight into the change in prevalence of marine debris entanglement over time. To address these, a mixed methods approach was employed, involving a literature review and an elicitation of expert opinions. Given the difficulty of acquiring robust standardised data, this review is intended to highlight the value of mixed methods as a first step to understand complex conservation issues, and to provide suggestive yet relevant indications as to the scale of the threat of entanglement to marine turtles.

MATERIALS AND METHODS

Literature review

In January 2016 and again in June 2017 (during the manuscript review process), all relevant literature was reviewed that may have contained records of marine turtle entanglement. ISI Web of Knowledge, Google Scholar and the Marine Turtle Newsletter (www.seaturtle.org) were searched for the terms 'entanglement', 'entrapment', 'ensnare' or 'ghost fishing' and 'turtle'. The first 200 results were viewed, with results very rarely fulfilling the criteria after the first 20; spurious hits were ignored and all relevant references were recorded and investigated.

Elicitation of expert opinions

During the period 1–30 April 2016, an online questionnaire survey was conducted to investigate 3 main topics of interest: (1) the occurrence and global spatial distribution of sea turtle entanglement; (2) species, life stage and debris type involved; and (3) the change in entanglement prevalence over time. A total of 20 questions requiring both open and closed responses from a range of experts were used to obtain insight into the scale of marine turtle entanglement. We clearly explained to the respondents the definition of 'marine turtle entanglement' specific to this study. Grid-like responses and Likert scales, offering potential answers from a range of ordinal options, were used to aid in achieving a quantitative assessment of the issues (Elaine & Seaman 2007) (see Box S1 in the Supplement at www.int-res.com/articles/suppl/n034p431_supp.pdf).

Potential participants for this questionnaire were identified from lead authorship of papers compiled in the recent review on the effects of marine plastic debris on turtles from Nelms et al. (2016), and our review due to their involvement in research into marine debris. From reviewing the few published reports, it was apparent that governmental stranding networks, sea turtle rescue and rehabilitation centres and conservation projects may also hold many unpublished records of entanglement occurrence. A comprehensive list of such organisations from seaturtle.org (www.seaturtle.org/groups/; accessed 24 March 2016) was used to find more expert contacts to participate in the questionnaire. Additionally, considering the aim of attaining an appropriate number of respondents while avoiding potential sampling biases due to researchers' personal networks and perceptions about the issue (Newing 2011), we employed respondent-driven sampling; this purposive sampling approach involves requesting those directly contacted to recruit additional participants among colleagues, peers and other organisations that may have knowledge of additional records of marine turtle entanglement.

From this first questionnaire, an initial report was produced and sent to the expert respondents (n = 106) to share the results and thoughts that arose from the first questionnaire. This included 8 initial figures produced from the data given by respondents in the original questionnaire to aid feedback of our results (these were draft versions of Figs. 2, 3 & 4). Following this, during the period 24 May to 30 June 2016, a follow-up questionnaire survey was conducted with the expert participants of the first questionnaire survey who were then invited to comment and answer 10 open and closed questions (see Box S2 in the Supplement). This aimed to further understand the challenges, future requirements (both research and priority actions) and perceptions of the likelihood of population level effects of marine turtle entanglement. In this second questionnaire, respondents were asked to comment on our initial results and to provide suggestions on future knowledge gains and actions. Their answers were categorised using an inductive approach; summary themes were identified through the process of directly examining the data (Elo & Kyngäs 2008), instead of having predefined categories.

RESULTS

Literature review

Our literature search yielded 23 reports regarding entanglement in multiple species of marine turtles, the majority of which were peer-reviewed publica-

tions (n = 17) with additional grey literature reports (n = 6). Species included loggerhead *Caretta caretta* (n = 7), green *Chelonia mydas* (n = 7), leatherback *Dermochelys coriacea* (n = 5), hawksbill *Eretmochelys imbricata* (n = 5), olive ridley *Lepidochelys olivacea* (n = 9) and flatback *Natator depressus* (n = 2). There were no records for Kemp's ridley *Lepidochelys kempii* (Table 1). Of these publications, 18 reported entanglement due to ghost fishing or fisheries materials and 7 recorded entanglement in land-based plastic debris; 7 publications reported the size range and life stage of the entangled turtles. These publications highlighted a range of impacts of entanglement, such as serious wounds leading to maiming, amputation or death, increased drag, restricted movement or choking that were further illustrated by photographs from collaborating experts (Fig. 1).

Elicitation of expert opinions

Survey response rates and demographics

From an estimated pool of ca. 500 potential contacts, the 'Marine Turtle Entanglement Questionnaire' was received and completed by a total of 106 expert respondents from 43 countries. However, due to the anonymous nature of the survey and the potential augmentation from the use of respondent-driven sampling, it is not possible to determine how many of those initially contacted took part in the survey. All ocean basins were covered; the respondents' main oceanic region of work was given as: Atlantic (34.8%; n = 39), Pacific (18.9%; n = 20), Caribbean (25.5%; n = 27), Mediterranean (9.4%; n = 10) and Indian (9.4%; n = 10). Respondents experienced a wide

Fig. 1. Impacts of marine turtle entanglement: (a) live leatherback turtle entangled in fishing ropes which increases drag, Grenada 2014 (photo: Kate Charles, Ocean Spirits); (b) drowned green turtle entangled in ghost nets in Uruguay (photo: Karumbe); (c) live hawksbill turtle entangled in fishing material constricting shell growth, Kaeyama Island, Japan 2001 (photo: Sea Turtle Association of Japan); (d) live hawksbill turtle with anthropogenic debris wrapped around front left flipper constricting usage of limb which could lead to amputation and infection, Kaeyama Island, Japan 2015 (photo: Sea Turtle Association of Japan). All photos used with express permission

Table 1. Summary of all studies on entanglement of marine turtles in plastic debris. CCL: curved carapace length (cm); na: not available

Ocean basin/ Species	Study area	Reference	Year of study	N	CCL range	Pelagic juvenile	Neritic juvenile	Adult	Debris type
Loggerhead turtle *Caretta caretta*									
Atlantic Ocean	Northeastern (Boa Vista, Cape Verde Islands)	López-Jurado et al. (2003)	2001	10	62.0–89.0	X	✓	✓	Fishing
	Northeastern (Terceira Island, Azores)	Barreiros & Raykov (2014)	2004 –2008	3	37.3–64.1	X	✓	✓	Fishing/land-based
	Northeastern (Gran Canaria, Spain)	Orós et al. (2016)	1998–2014	945	Unknown	✓	✓	✓	Fishing/land-based
Mediterranean Sea	Tyrrhenian sea (Island of Panarea, Sicily)	Bentivegna (1995)	1994	1	48.5	X	✓	X	Land-based
	Central Mediterranean (Italy)	Casale et al. (2010)	1980–2008	226	3.8–97.0	✓	✓	✓	Fishing/land-based
	South Tyrrhenian sea	Blasi & Mattei (2017)	2009–2013	5	Unknown	na	na	na	Fishing/land-based
Global		Balazs (1985)	1967–1984	5	Unknown	✓	✓	✓	Fishing
Green turtle *Chelonia mydas*									
Indian Ocean	North (Maldives)	Stelfox & Hudgins (2015)	2013–2015	2	Unknown	na	na	na	Fishing
	Northeastern (Darwin, Australia)	Chatto (1995)	1994	1	35	X	✓	X	Fishing
	Northeastern (Australia)	Wilcox et al. (2013)	2005–2009	14	Unknown	na	na	na	Fishing
Global		Balazs (1985)	1967–1984	24	Unknown	✓	✓	✓	Fishing (21), land-based (3)
Pacific Ocean	Central (Hawaii)	Francke et al. (2014)	2013–2014	51	Unknown	✓	✓	✓	Fishing
		Chaloupka et al. (2008)	1982–2003	43	20.0–100.0	✓	✓	✓	Fishing
Caribbean Sea	Southeastern (Venezuela)	Barrios-Garrido et al. (2013)	2013	1	Unknown	na	na	na	Fishing
Leatherback turtle *Dermochelys coriacea*									
Indian Ocean	North (Maldives)	Stelfox & Hudgins (2015)	2013–2015	1	Unknown	na	na	na	Fishing
Pacific Ocean	Northeastern (USA)	Moore et al. (2009)	2001–2005	1	Unknown	na	na	na	Fishing
Atlantic Ocean	Northwestern (USA)	Hunt et al. (2016)	2007–2013	8	Unknown	na	na	na	Fishing
	Northwestern (USA)	Innis et al. (2010)	2007–2008	7	Unknown	na	na	na	Fishing
Global		Balazs (1985)	1967–1984	5	Unknown	X	X	✓	Fishing
Hawksbill turtle *Eretmochelys imbricata*									
Indian Ocean	North (Maldives)	Stelfox & Hudgins (2015)	2013–2015	6	Unknown	X	✓	X	Fishing
	Northeastern (Darwin, Australia)	Chatto (1995)	1994	1	32.5	X	✓	X	Fishing
	Northeastern (Australia)	Wilcox et al. (2013)	2005–2009	35	Unknown	na	na	na	Fishing
	Northeastern (Northern Territory, Australia)	White (2006)	2004	2	Unknown	X	✓	X	Fishing
Global		Balazs (1985)	1967–1984	9	Unknown	✓	✓	✓	Fishing (8), land-based (1)
Olive ridley turtle *Lepidochelys olivacea*									
Indian Ocean	North (Maldives)	Anderson et al. (2009)	1998–2007	25	10.0–61.0	✓	✓	X	Fishing (22), land-based (3)
	North (Maldives)	Stelfox & Hudgins (2015)	2013–2015	163	Unknown	✓	✓	✓	Fishing
	Northeastern (McCluer Island, Australia)	Jensen et al. (2013)	Unknown	44	Unknown	na	na	na	Fishing
	Northeastern (Australia)	Wilcox et al. (2013)	2005–2009	53	Unknown	na	na	na	Fishing
	Northeastern (Darwin, Australia)	Chatto (1995)	1994	2	64	X	X	✓	Fishing
	Northwestern (Seychelles)	Remie & Mortimer (2007)	2007	1	Unknown	X	✓	X	Unspecified
Atlantic Ocean	Southwestern (Brazil)	Santos et al. (2012)	1996–2011	18	2.01–80.0	X	✓	✓	Fishing
Global		Balazs (1985)	1967–1984	7	Unknown	✓	✓	✓	Fishing
Pacific Ocean	Central (Hawaii)	Francke et al. (2014)	2013–2014	1	Unknown	na	na	na	Fishing
Flatback turtle *Natator depressus*									
Indian Ocean	Northeastern (Darwin, Australia)	Chatto (1995)	1994	1	25.5	X	✓	X	Land-based
	Northeastern (Australia)	Wilcox et al. (2013)	2005–2009	3	Unknown	na	na	na	Fishing
Multiple									
Indian Ocean	Northeastern (Australia)	Wilcox et al. (2015)	2005–2012	336	Unknown	na	na	na	Fishing
Pacific Ocean	Southwestern (Australia)	Meager & Limpus (2012)	2011	5	Unknown	na	na	na	Fishing

range in the number of annual stranding cases in their respective study sites (annual maxima given in the survey; mean ± SE = 239.9 ± 71.7, range = 0 to 4100, n = 97) but in total, through addition of the respondents' answers, they are responsible for attending an estimated 23 000 stranded turtles yr^{-1}. Respondents also generally had many years of experience dealing with and reporting marine turtle strandings (range = 2 to 42 yr, mean ± SE = 15.6 ± 1.1, n = 98), confirming them as having relevant experience to answer the survey. The second follow-up questionnaire sent to all respondents (n = 106) received 63 responses with respondents from 31 countries.

Rates of entanglement

A majority of respondents (84.3%; n = 101) had encountered cases in which turtles were entangled in anthropogenic debris. When broken down by species, the proportion of stranded turtles that were entangled did not differ significantly (Kruskal-Wallis: $\chi^2 = 4.59$, df = 6, p = 0.59) (Fig. 2a). There was a low percentage incidence for all species, with the grand median rate of 5.5%, although there was considerable inter- and intraspecific variation, with incidences in different responses ranging from 0 to 95.5%. In terms of the proportion of marine turtles alive when found entangled, there were significant interspecific differences (Kruskal-Wallis: $\chi^2 = 19.62$, df = 6, p = 0.003). The proportion found alive (grand median = 9.4%) was significantly higher in green (25.5%) and loggerhead (15.5%) turtles than in all other species (5.5%) (Fig. 2b).

Entanglement rates also differed amongst life stages for each species. Whilst respondents indicated that all life stages of each species had been affected by entanglement, the results suggested adults were most impacted in leatherback and olive ridley turtles, whereas for the remaining species respondents indicated a higher rate of entanglement in juveniles (pelagic and neritic; Fig. 3).

When considering this issue over time (over the last 10 yr), a similar proportion of respondents (35.8% of 106) thought the prevalence of entanglement had increased or remained the same, while the remainder thought it had decreased (8.5%) or were unsure (19.8%). Among those respondents that noted an increase, some (n = 4) suggested that this may be caused by an increase in reporting and awareness, while others (n = 9) indicated the development of coastal fishing activities might be a factor. When asked to consider a shorter time period (the last 5 yr),

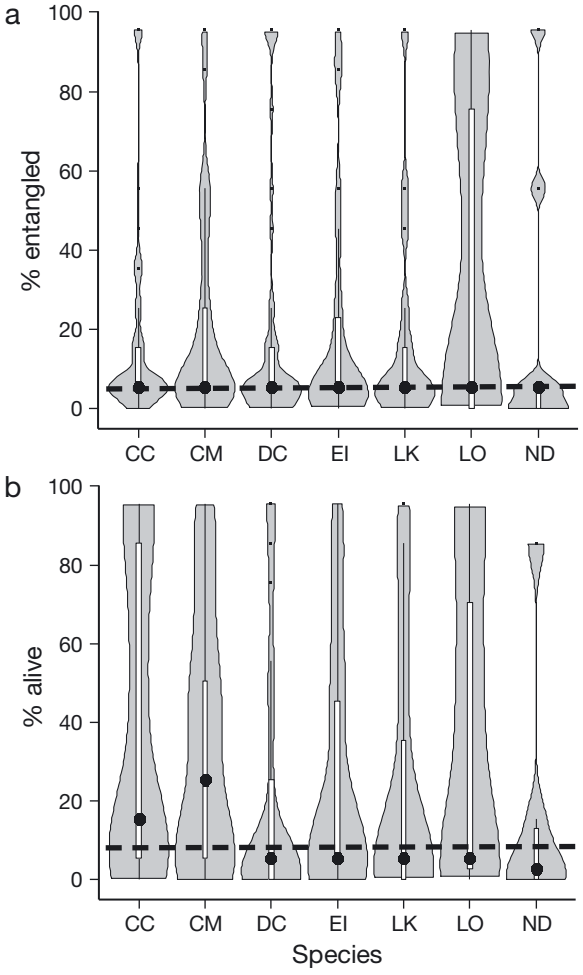


Fig. 2. Inter-species comparison of the proportion of: (a) stranded individuals found entangled and (b) individuals found alive when discovered entangled. Violin plots show the kernel density of data at different values. Median (black dot) with interquartile range boxplot (black/white) and grand median (black dashed line). Turtle species abbreviations: CC: loggerhead *Caretta caretta*; CM: green *Chelonia mydas*; DC: leatherback *Dermochelys coriacea*; EI: hawksbill *Eretmochelys imbricata*; LK: Kemp's ridley *Lepidochelys kempii*; LO: olive ridley *Lepidochelys olivacea*; ND: flatback turtle *Natator depressus*

the majority of respondents believed that the prevalence of entanglement they had experienced had remained stable (51.9%), whilst the others thought it had increased (29.2%), decreased (3.8%) or were not sure (15.1%).

Entanglement materials

The majority of entanglements recorded were with lost/discarded fishing gear (Fig. 4). A clear distinction was made between 'active' and 'lost/discarded'

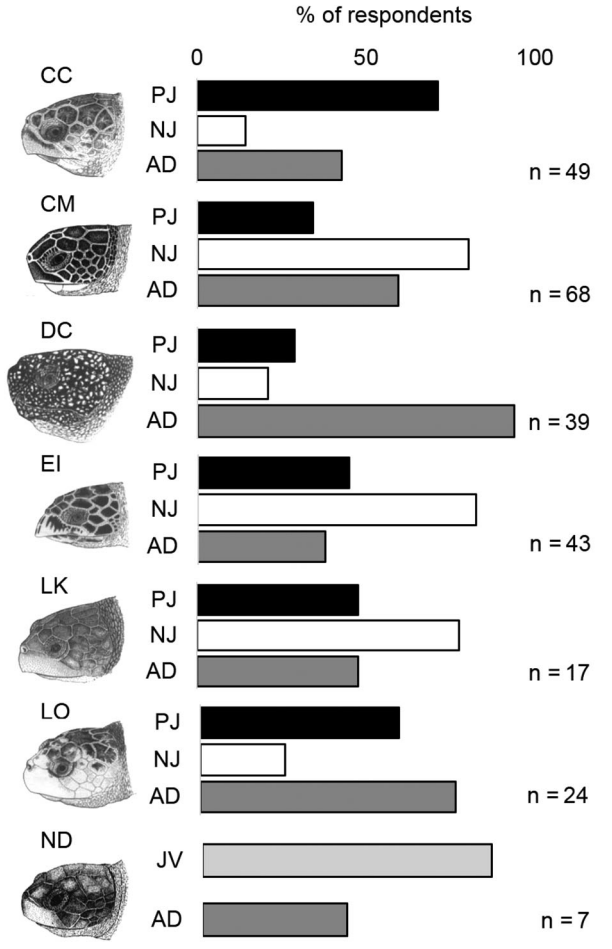


Fig. 3. Inter-specific comparison of the breakdown of entangled sea turtle species by life stage. Black: pelagic juveniles (PJ); white: neritic juveniles (NJ); light grey: juveniles (JV); dark grey: adults (AD); see Fig. 2 for species abbreviations. Flatback turtles were only categorised into juvenile or adult classes with advice from species experts. Sea turtle skull figures used with permission of WIDECAST; original artwork by Tom McFarland

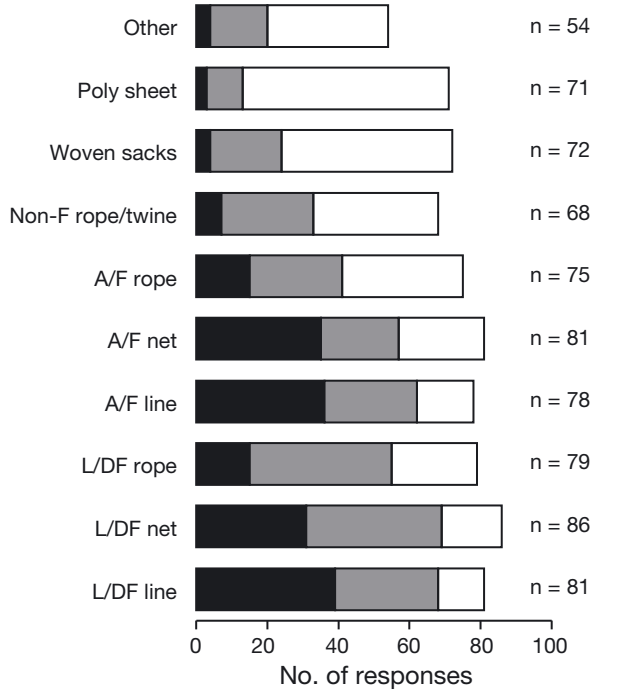


Fig. 4. Entangling materials. L/DF: lost/discarded fishing; A/F: active fishing; Non-F: non fishing; Poly sheet: polyethylene sheeting. Black: common (10% or more of cases); grey: sometimes (less than 10% of cases); white: never. Not all participants categorised each material; total number of responses for each material shown on the right of the graph

fishing gear to try and separate incidents due to bycatch and subsequent stranding from those caused by ghost fishing. The number of responses on the occurrence of ghost fishing (GF) through discarded fishing debris (rope, net and line) was generally slightly higher than for bycatch (BC) through active gear.

A smaller percentage of respondents specified cases of turtle entanglement in land-based sources, from polythene sheeting (n = 71), woven sacks (n = 72) and non-fishing rope/twine (n = 68). But in only a few incidences were these said to be common occurrences (polythene sheeting [n = 3], woven sacks [n = 4], non-fishing rope/twine [n = 7]). Respondents were asked to comment on the occurrence of 'other' entangling materials (n = 54) and to provide examples (n = 20) that caused turtle entanglement. This included debris from land-based sources (plastic-balloon string, canned drink '6-pack' rings, kite string, plastic chairs, plastic packaging straps, wooden crates and weather balloons) and debris from other maritime activities (boating mooring line, anchor line and discarded seismic cable).

Scale of issue

In order to obtain further insights into the potential scale of this issue, respondents to the second survey were asked whether they thought entanglement in anthropogenic debris is causing population-level effect in marine turtles. Of the 63 respondents, 84.1% thought that this was probable, very likely or definite (see Fig. S1 in the Supplement). There was no significant difference in scaled responses by ocean basin (Kruskal-Wallis: $\chi^2 = 1.82$, df = 4, p = 0.77). In order to assess the relative importance of different threats according to experts, we also sought the experts' opinions on how they thought entanglement in anthro-

pogenic debris compared to other threats to marine turtles (i.e. 'plastic ingestion', 'oil pollution', 'fisheries bycatch', 'direct exploitation' and 'climate change'). Although between 6.35 and 25.4% were unsure, there was a strong opinion that plastic ingestion and fisheries bycatch were greater threats, and that oil pollution, climate change and direct exploitation were less severe threats than entanglement (Fig. 5).

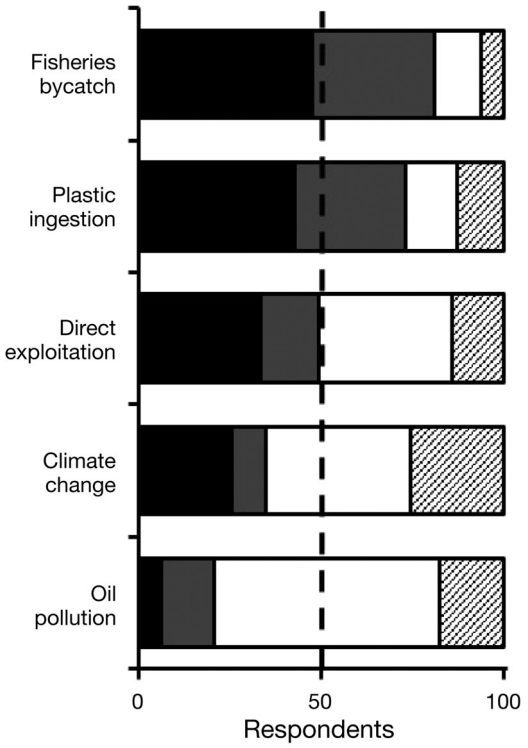


Fig. 5. Responses to comparison of other threats faced by marine turtles compared to entanglement (n = 63). Black: greater than entanglement; grey: similar threat; white: less than entanglement; striped: unsure

Challenges, priority actions and research needs

Respondents to the second survey converged on a limited number of themes when considering the challenges, research needs and priority actions within marine turtle entanglement. The challenges to addressing the issue (115 suggestions) could be grouped into 5 major categories: law and enforcement (23.5%; n = 27); sources and spatial extent of entanglement materials (24.3%; n = 28); education and innovation (24.3%; n = 28); understanding the full extent of the threat (18.3%; n = 21); and human response to entangled turtles (9.6%; n = 11) (Table 2). Seven major research areas were suggested by respondents (91 suggestions): more specific reporting and monitoring or a common database (23.1%; n = 21); mapping the threat/spatio-temporal hotspots (31.9%; n = 29); identifying entanglement materials and sea turtle interactions (24.2%; n = 22); understanding post-release mortality and physical effects (3.3%; n = 3); socio-economic impacts (4.4%; n = 4); innovation of new replacement materials (6.6%; n = 6); and demographic risk assessments (6.6%; n = 6) (Table 3). Priority actions (n = 121 suggestions) that respondents believe would help reduce turtle entanglement were grouped into 5 major areas: education/stakeholder engagement (31.4%; n = 38); fisheries management and monitoring (26.4%; n = 32); research (5%; n = 6); law and enforcement (20.7%; n = 25); and development of alternative materials and methods (16.5%; n = 20) (Table 4).

DISCUSSION

Global distribution

Our review and elicitation of expert opinions demonstrate that marine turtle entanglement is an issue operating at a global scale, occurring in all species, throughout their geographic range. We sought to answer key knowledge gaps surrounding the issue of turtle entanglement in marine debris as previously highlighted by Vegter et al. (2014) and Nelms et al. (2016). Difficulties in investigating these knowledge gaps are in part due to a lack of robust data. This highlights the importance of using mixed methods to access expert opinion to gain an insight into this global threat. The growing use of expert knowledge in conservation is driven by the need to identify and characterise issues under limited resource availability, and the urgency of conservation decisions (Martin et al. 2012).

Acknowledging the incomplete coverage of our estimates, given the mean estimated number of strandings and mortality rates, in the order of 1000 turtles die annually as a result of entanglement in the areas monitored by our respondents. These levels are likely a profound underestimation of the scale of this issue as the coverage of these actors is far from comprehensive. Second, it is well known that not all dead turtles strand (Epperly et al. 1996, Sasso & Epperly 2007), especially small and pelagic animals, and there can also be decay of entangled animals. Additionally, some of our respondents commented that detection of stranded animals may be further confounded due to take of stranded animals for human consumption.

Table 2. Summary of major challenges regarding marine turtle entanglement as listed by respondents

Challenge category	% of suggestions (n = 115)	Challenges described	Direct quotes from respondents
Law and enforcement	23.5	Management of both industrial and small-scale artisanal fisheries	'Under-resourced fisheries management of small-scale fisheries'
		The issue of discarded fishing gear at sea	'Trawlers should file a report anytime they lose netting'
		Ineffectiveness of Marine Protected Areas	'Shifting climate may render Marine Protected Areas as ineffective'
Source of entanglement materials and extent	24.3	Estimating the amount and durability of entangling material entering the sea	'Entangling material tends to be durable, so even if management scheme is put into place, have to deal with historic material already in the ocean'
		Retrieving lost fishing gear	'In my region, lost/discarded fishing lines are a big issue'
		Lack of accountability	'Inability to determine source of entanglement debris (no accountability)'
Education and innovation	24.3	Fisherman education and awareness	'Engagement/education/enticement to bring artisanal fishers in developing countries to a want to reduce turtle mortality' 'Figuring out how to reach out to boaters/fishermen with making them want to support sea turtle friendly habits'
		Developing a discipline to avoid abandonment of fishing gear	'Addressing amateur/recreational fishers is really hard. In my opinion, most of the discarded fishing lines are left by this group'
		Sourcing alternative materials	'Creation of degradable nylon'
Understanding the full extent of the threat	18.3	Lack of stranding networks' ability to measure the impact of the extent of the threats in multiple areas	'It is hard to estimate the total amount of entangled turtles, since these animals are highly migratory and tend to be scattered over wide areas. Additionally turtles that become entangled may quickly die and be predated. Scavengers, predators, wind and currents may prevent carcasses from coming ashore' 'Most entanglement records rely on land-based sampling and stranding do not represent total deaths at sea' 'It is hard to distinguish marine debris from active and ghost fishing gears'
		Difficulty in determining if entanglement occurred pre- or post-mortem	'Difficulty in determining if entanglement occurred pre- or post-mortem (for some entanglement types, such as discarded nets/line)'
		Survivorship of turtle found entangled alive	'Limited post-release monitoring of live entangled turtles'
Response to entangled turtles	9.6	Detangle permits	'Very few people are trained and permitted to disentangle them'
		Discovery times need to be quick	'Discovering entangled turtles quickly' 'Entangled turtle can be challenging to disentangle especially if they are not anchored and instead are free swimming'
		Ineffectiveness of reporting systems	'Having a good system in place that stranding will be reported (people that see an entangled turtle have to be able to notify the correct organization)'
		Lack of rehabilitation resources for entanglement incidents	'Lack of rehabilitation resources for turtles hurt in incidents of entanglement'

Table 3. Summary of research needs regarding marine turtle entanglement as listed by respondents

Research need category	% of suggestions (n = 91)	Research needs described	Direct quotes from respondents
More specific reporting and monitoring/common database	23.1	Creation of a common database	'A common database, long lasting surveys and a programme on a national base for monitoring of the state of debris in the sea'
		An increase in specificity of reporting of entanglement cases	'Better monitoring/reporting of entanglement cases by species, life stage, region'
			'Establish a protocol for sea turtle strandings networks for identify entanglements and report these'
		Collaboration of resource users in the marine environment	'More collaboration with resource users in the marine environment in respect to reporting cases of entanglement'
			'Getting information from fishermen when turtles get entangled. Support to Fisheries Division who can provide accurate information on net damage from reports by fishermen. Only a small percentage of stranded turtles will wash up ... carcasses may become destroyed prior to reaching those coasts'
Mapping the threat/spatio-temporal hotspots	31.9	Using stakeholder knowledge	'Surveys to fishermen (industrial, artisanal and sport) to understand where and when they discard nets or lines and in water monitoring programs in coastal areas with high pressure of artisanal and sport fishing'
		Identifying and mapping the entanglement rates due to different gear types and materials	'Understanding where the event occurs, such as targeting if the problem is more from floating debris versus debris in water column'
		Modelling/mapping patterns of debris distribution, patterns of marine turtle migrations and the characterization of fisheries distributions	'Understanding overlap between sea turtle habitats (e.g. nesting and feeding grounds) with areas of high debris concentration (e.g. convergence zones)'
			'Spatio-temporal scales. Hotspots'
Entanglement materials and sea turtle interactions	24.2	Studying sea turtle and debris behaviour and their interactions	'Behavioural (foraging or sheltering) traits in different turtle species or populations that may them more vulnerable to entanglement'
			'Investigate the behavioural characteristics of the turtles that lead to their entrapment in fishing gear with a view to improving mitigation actions'
Post-release mortality and survival/physical effects	3.3	Understanding true post-release mortality and morbidity	'The effects of flipper amputations on survival'
Socio-economic impact	4.4	Special focus on the fisher community	'What are the opportunities and barriers to intervention?'
Innovation of new replacement materials & methods	6.6	Innovation of biodegradable alternatives to commonly used plastic materials	'Alternative materials for fishing and other things/activities'
Demographic risk assessments	6.6	Development of demographic risk assessments for threatened populations of turtles	'Develop the appropriate population demographic models for marine turtles to allow for assessment/identification of those mortality factor that are not detrimental to maintaining robust non threatened population of turtle'

Table 4. Summary of priority actions regarding marine turtle entanglement as listed by respondents

Priority actions category	% of suggestions (n = 121)	Priority actions described	Direct quotes from respondents
Education/stakeholder engagement	31.4	Fisher involvement/education	'Develop questionnaire for fishermen for their recommendations on how it would be possible to reduce turtle entanglement'
			'Partnership with local fishermen to locate and remove abandoned or lost fishing gear (ghost gear). Financial incentives to return discarded gears to shore'
		Community/public awareness campaigns on marine litter	'Organizing campaigns with scuba divers to clean sea bottom from the man debris and ghost nets/discarded fishing lines'
			'Implement an environmental stewardship certificate system among ocean users and create a global open access database of entanglements to facilitate research efforts'
Fisheries management and monitoring	26.4	The development of traceable gear	'Developing/using traceable gear in combination with introducing a fining policy'
		Stricter regulations	'Increased collaborations with commercial fisherman and recreational fisherman to better understand their needs and the needs of the turtles....and how these can be combined'
Research/knowledge	5	The implementation of the research needs stated in Table 3	'We cannot say before understanding the main reasons, main sources and main habitats or localities in which entanglement occurs'
Law and enforcement on entanglement material	20.7	Banning at-sea disposal of entangling materials	'Enforcement of laws banning at-sea disposal of entangling material'
		Better waste management and increased recycling efforts	'Reduction of manmade debris, better waste management, more biodegradable products'
Development of alternative materials/methods	16.5	Development of alternative materials/ methods	'Development of less environmentally persistent materials to be used in nets, fishing line, etc.'
		Shifting gear type/increasing the use of biodegradable materials	'Different strategies to different fishing gear; from the coastal sport fishermen to high seas industrial fishermen'
			'Introduce biodegradable chord into selected net fisheries with high loss to ghost nets'

Species differences

Although there was no interspecific difference in the incidence of entanglement, most peer-reviewed publications featured olive ridley turtles, with some experts reporting high incidences of entanglement for this species. Stelfox et al. (2016) noted that olive ridley turtles accounted for the majority of sea turtles identified as entangled (68%; n = 303), and this could be for the following reasons. Firstly, this species, which often exhibits mass nesting in the hundreds of thousands of individuals, is highly numerous, and at particularly high densities in some areas, leading to entanglement hotspots (Jensen et al. 2006, Koch et al. 2006, Wallace et al. 2010a). Secondly, the olive ridley forages along major oceanic fronts which are known to aggregate marine debris (Polovina et al. 2004, McMahon et al. 2007). Finally, their generalist feeding behaviour potentially attracts them to feed opportunistically on biofouled marine debris such as ghost gear (Stelfox et al. 2016).

Life stages

Entanglement was reported to occur in all life stages (pelagic juveniles, neritic juveniles and adults) across all species (the exception being flatback turtles which have no pelagic juveniles; Hamann et al. 2011). Perhaps of greatest concern is the signal of high entanglement incidence in the pelagic juvenile stage: despite the general inaccessibility of sampling this life stage, they are still appearing as stranded entangled. The currents that transport hatchlings to oceanic convergence zones are also now recognised as concentrating floating anthropogenic debris, creating the capacity for an ecological trap for these young turtles, whether it be through ingestion or entanglement (Nelms et al. 2016, Ryan et al. 2016). Many respondents considered that entanglement could be having a population level effect; a distinct possibility if this there is a large impact on this cryptic life stage and on pelagic foraging adults (Mazaris et al. 2005).

Entangling materials

Respondent data highlighted that the majority of entanglements were the result of fishery-based material and other maritime activities. The issue of ghost fishing featured highly, with numerous responses reporting entanglement within lost/discarded gear. This gear is often lost, abandoned or discarded when it becomes derelict, attracting scavengers and acting as FADs (Gilman 2011). Subsequently, species such as marine turtles become entangled within the gear, perhaps encouraged by this process of 'self-baiting' (Matsuoka et al. 2005).

Change in fishing practice

The issue of ghost fishing appears to have worsened since the 1950s, as the world's fishing industries have replaced their gear, which was originally made of natural fibres such as cotton, jute and hemp, with synthetic plastic materials such as nylon, polyethylene and polypropylene. Manufactured to be resistant to degradation in water means that once lost, it can remain in the marine environment for decades (Good et al. 2010). Furthermore, there has also been a shift in the type of synthetic nets being selected; for example, fishers in part of Southeast Asia now increasingly favour superfine nets. Although this can help increase catches, the twine thinness means that they break easily and are difficult to repair once damaged (Stelfox et al. 2016). The incidences of entanglement caused by this form of pollution in our expert surveys indicates that this source of mortality for marine turtles mirrors that in marine mammals and sea birds, which has increased substantially over the last century (Tasker et al. 2000, Good et al. 2010, McIntosh et al. 2015).

Differentiation from bycatch

It is quite plausible that ghost fishing may be working synergistically alongside bycatch, but because of its more cryptic nature this means that understanding its role in marine turtle mortality is much more difficult. Bycatch is better understood. For example, the analysis of catch rates in the Mediterranean allowed for the estimation of 132 000 captures and 44 000 incidental deaths per year (Casale 2011). Likewise, cumulative analysis of catch rates in US fisheries estimated a total of 71 000 annual deaths prior to the establishment of bycatch mitigation methods. Since these measures were implemented, mortality estimates are ~94% lower (4600 deaths yr^{-1}) (Finkbeiner et al. 2011). This highlights the importance of informed estimates to monitor the success of mitigation methods. In addition to bycatch mortality estimates, spatial and temporal patterns of bycatch inci-

dences can be identified. Using onboard observer data, Gardner et al. (2008) found seasonal changes in catch distributions of loggerhead and leatherback turtles in the North Atlantic, with patterns of spatial clustering from July to October. Analysed on a global scale, Wallace et al. (2010b) were able to highlight region–gear combinations requiring urgent action such as gillnets, longlines and trawls in the Mediterranean Sea and eastern Pacific Ocean. Generating such estimates of catch rates and spatial/temporal patterns for entanglement are not yet possible due to the lack of quantitative information.

Land-based plastic entanglements

The domination of fisheries-based materials in the results does not mean that land-based plastics are not a source of entanglement. The increased input of plastic debris from terrestrial run-off means that these interactions are only likely to increase (Jambeck et al. 2015). Our literature search and 'other' materials stated by respondents contained a variety of items causing entanglement that could be decreased by reduction of use, replacement with more degradable alternatives and better waste management and recycling. The prevalence of these materials in the marine environment will very much depend on future waste governance, especially in those countries that generate the most plastic waste (Jambeck et al. 2015). A future technological solution which is currently being investigated or adopted in high plastic-generating countries such as Thailand and India is the pyrolysis of plastics. This process produces fuel from waste plastic, a better alternative to landfill and a partial replacement of depleting fossil fuels (Wong et al. 2015).

Caveats

It is important to recognise the biases associated with using stranding animals for data collection. Within and between stranding sites there are differences in turtle foraging ecology, life stages and proximity to human habitation (Bolten 2003, Rees et al. 2010), and therefore they are exposed to different levels and types of potential entangling materials. Individual turtles therefore may not represent a homogeneous group in terms of entanglement occurrence within that population (Casale et al. 2016). Additionally, recovered carcasses represent an unknown fraction of at-sea mortalities, with physical oceanography (e.g. currents) and biological factors (e.g. decomposition) affecting the probability and location of carcass strandings (Hart et al. 2006). However, examining reports of stranded animals represents a vital opportunity for research and can provide insights into the impacts of anthropogenic threats which would otherwise go undetected (Chaloupka et al. 2008, Casale et al. 2010). In addition, stranding information aids with the assessment of harder-to-access life stages, yielding key information on the risk to specific resident populations and contributing to building a worldwide perspective for conservation issues (Chaloupka et al. 2008, Casale et al. 2016). Indeed, this was the aim of our study: using stranding data from expert respondents to gain an initial indication of the estimated magnitude of this threat.

Surveying experts can be a powerful tool for obtaining insights on particular topics not widely known by others (Martin et al. 2012). Expert knowledge and opinions may be the result of training, research, skills and personal experience (Burgman et al. 2011a). In this study, we sought the opinions of conservation scientists and practitioners with experience in marine turtle entanglement and strandings. Due to the purposive sampling nature of our approach, we aimed to identify people with relevant experiences instead of focusing on obtaining a random selection of representatives; this is a widely used practice when undertaking social surveys that focus on particular subgroups or specialists (Newing 2011). Nevertheless, expert knowledge and opinions are also known to be subject to biases, including overconfidence, accessibility and motivation (see e.g. Burgman et al. 2011b and Martin et al. 2012). In the absence of empirical data to validate our findings, this remains as simply suggestive but nevertheless relevant information in terms of identifying a potentially important conservation issue and providing relative indications of the scale of entanglement as a threat to sea turtles.

Future actions and recommendations

Ghost fishing

Issue and policy. Presently, a large knowledge gap exists regarding effects of ghost fishing. While there has been some progress in documenting the frequency of loss from passive gear such as gillnets, little is known about loss from active gears; effective methodology to estimate the persistence of types of

gear such as trawl nets has yet to be developed (Gilman et al. 2013). While it would be optimal to switch all gear to more biodegradable materials, synthetic materials will continue to be used within fisheries for the foreseeable future. This is an issue that has been highlighted in policy by the Food and Agriculture Organization (FAO), who recommend the identification, quantification and reduction of mortality caused by ghost fishing by implementing this into fisheries management plans, increasing scientific information and developing mitigation strategies; but this appears still to be in its infancy (Gilman et al. 2013). This is also reflected in mandates within the International Maritime Organisation (IMO) and International Convention for Prevention of Pollution from Ships (MARPOL Annex V) (Stelfox et al. 2016).

Need for a global database and spatial hotspot identification. Undoubtedly a common global metadatabase recording the spatial distribution and abundance of possible entangling ghost gear as well as incidences of marine turtle entanglement incorporating a unit of effort metric would assist in quantifying the mortality due to ghost gear that is needed to inform policy (Nelms et al. 2016). A recent global review (dominated by the Atlantic and Pacific oceans) on marine megafauna by Stelfox et al. (2016) reported a total of 5400 individuals of 40 species that had been associated with ghost gear between 1997 and 2015. They suggested this was a great underestimate due to lack of capacity to record incidence. Such data could feed into one of the major research priorities emphasised by respondents; modelling spatio-temporal hotspots of entanglement. An innovative study by Wilcox et al. (2013) used beach clean data and models of ocean drift to map the spatial degree of threat posed by ghost nets for marine turtles in northern Australia and map areas of high risk. With the input of more specific marine location data on ghost gear and the advocacy of the use of ever improving modelling, this could provide a powerful tool in the future.

Education and stakeholder engagement

Local initiative to reduce debris causing entanglement. On a more local and regional scale, many initiatives are being brought into place to encourage a reduction in the amount of ghost gear/plastic debris entering the ocean and combat discarding at sea by working closely with community education and engagement; another highlighted topic by our respondents. There are numerous examples: the sea turtle conservation program in Bonaire has started a 'Fishing Line Project' (www.bonaireturtles.org/wpp/what-we-do/fishing-line-project) working with volunteers to train them on how to remove discarded line and nets from coral reefs, and the Zoological Society of London's 'Net-works' (www.net-works.com) initiative has established a supply chain for discarded fishing nets from artisanal fishing communities in the Philippines to a carpet manufacturing company. With further replication of such community-based projects and stakeholder engagement, especially with artisanal fisheries awareness, the potential exists to start targeting hotspots of marine vertebrate entanglement directly.

Stranding networks training. Another set of stakeholders which will be important to engage are stranding networks. Responses to entangled turtles can often be slow, and respondents commented that many are not trained in the correct protocols to safely remove entangling materials. If stranding networks were fully trained in a standardised protocol for removal, the techniques could then be passed on through educational training programmes to the fishing community, quickening the response to such incidences. This is already beginning to happen for bycatch cases; Sicilian fisherman now actively volunteer to take part in the rescue of turtles in difficulty and are trained in contacting the competent authorities for the transfer of turtles to the nearest recovery centres. This level of involvement by workers in the fishery sector was stressed and encouraged through both effective education activity and specific targeted study campaigns (Russo et al. 2014).

Future research avenues into marine turtle entanglement

Respondents raised the issue of post-release mortality and the importance of behavioural research into the interactions between marine turtles and potential entangling materials present in the marine environment. The prominence of this has been emphasised within other taxa; for example, post-release mortality can result from long-term chronic effects of injuries in pinnipeds even after the entanglement has been removed (McIntosh et al. 2015). Furthermore, it has been argued that some colonial seabirds released from entangling plastic would not survive without human intervention (Votier et al. 2011).

To validate the success of release protocols after entanglement incidents (as mentioned above), techniques could be employed from other areas of marine turtle research. Satellite telemetry has already been used in a multitude of ways to provide information on conservation issues facing marine turtles; a number of studies have used this technique to consider post-release mortality after bycatch fisheries interactions (reviewed in Jeffers & Godley 2016). Deploying tagged turtles that have been involved in entanglements could aid in the understanding of survival after these events as well as simultaneously providing information on the location of sea turtles, feeding into information on entanglement hotspots to target mitigation actions. The benefits of utilising such techniques have been illustrated in other endangered species facing entanglement, such as studying mortality of silky sharks *Carcharhinus falciformis* in the Indian Ocean; estimates derived from satellite tracking showed that mortality due to entanglement was 5 to 10 times that of known bycatch mortality and provided evidence for a call advising immediate management intervention (Filmalter et al. 2013).

Other research methods and ideas could be modified from the study of plastic debris ingestion by sea turtles. Studies are currently underway to understand the selective mechanisms that lead to ingestion of plastic pieces (Schuyler et al. 2014, Nelms et al. 2016). For instance, a study by Santos et al. (2016) used Thayer's law of countershading to assess differences in the conspicuousness of plastic debris to infer the likelihood that visual foragers (sea turtles) would detect and possibly ingest the plastic fragments. Similar studies could be conducted to comprehend the underlying behavioural and physiological mechanisms that influence turtles to approach potential entangling materials when encountering them within the marine environment.

Similarly, comprehending how important the level of biofouling on this synthetic debris is in contributing to the likelihood of entanglement will be important. Total fish catches by monofilament gillnets in Turkey was lower, as a result of accumulating detritus and biofouling increasing the visibility of the nets in the water column (Ayaz et al. 2006). Furthermore, the level of biofouling could indicate the age of ghost gear entangling marine turtles. Retrieved lost/discarded fishing gears are usually found fouled by macro-benthic organisms, so if a relationship between soak time and biofouling level could be established, these organisms could provide a valid methodology to age the gear and enable better estimates of 'catches' made by the respective net (Saldanha et al. 2003).

Finally, it will be important to undertake demographic studies, calculating rates of entanglement, especially for specific populations that are known to be particularly vulnerable to a combination of other anthropogenic threats. For species such as pinnipeds, which are less elusive (hauling out on land) than marine turtles, the literature describes different methods. For example, a proportion derived from a count of entangled individuals from a sub-sample or an estimate of the total population (Raum-Suryan et al. 2009, McIntosh et al. 2015), or more recently, the use of mixed-effects models to obtain a prediction of the total number of seals entangled per year, by examining changes in entanglement rates over time and the potential drivers of these detected trends (McIntosh et al. 2015). However, this can only be achieved if reporting and recording such incidences in marine turtles improves in efficacy and standardisation.

CONCLUSIONS

Further research may show that the issue is more one of animal welfare than of substantive conservation concern to many marine turtle populations. It is clear, however, that entanglement with anthropogenic plastic materials such as discarded fishing gear and land-based sources is an under-reported and under-researched threat to marine turtles. Collaboration among stakeholder groups such as strandings networks, fisheries and the scientific community will aid in providing mitigating actions by targeting the issue of ghost fishing, engaging in education and producing urgently needed research to fill knowledge gaps.

Acknowledgements. The authors thank all respondents of the questionnaires for their invaluable knowledge and insights regarding this issue. We are grateful to Karen Eckert of WIDECAST for granting access to turtle graphics. E.M.D. received generous support from Roger de Freitas, the Sea Life Trust and the University of Exeter. B.J.G. and A.C.B. received support from NERC and the Darwin Initiative, and B.J.G. and P.K.L. were funded by a University of Exeter — Plymouth Marine Laboratory collaboration award which supported E.M.D. We acknowledge funding to T.S.G. from the EU Seventh Framework Programme under Grant Agreement 308370, and P.K.L. and T.S.G. received funding from a NERC Discovery Grant (NE/L007010/1). This work was approved by the University of Exeter, CLES ethics committee (Ref. 2017/1572). The manuscript was greatly improved by the input of the editor and 2 anonymous reviewers.

LITERATURE CITED

Anderson RC, Zahir H, Jauharee R, Sakamoto T, Sakamoto I, Johnson G (2009) Entanglement of live ridley turtles *Lepidochelys olivacea* in ghost nets in the equatorial Indian Ocean. Presented at the fifth session of the Indian Ocean Tuna Commission (IOTC) Working Party on Ecosystems and Bycatch, 12–14 October 2009, Mombasa

Ayaz A, Acarli D, Altinagac U, Ozekinci U, Kara A, Ozen O (2006) Ghost fishing by monofilament and multifilament gillnets in Izmir Bay, Turkey. Fish Res 79:267–271

Balazs GH (1985) Impact of ocean debris on marine turtles: entanglement and ingestion. In: Shomura RS, Yoshida HO (eds) Proceedings of the workshop on the fate and impact of marine debris, 27–29 November 1984, Honolulu, HI. NOAA Tech Memo NMFS-SWFC-54. US Department of Commerce, Washington, DC, p 387–429

Barnes DKA, Galgani F, Thompson RC, Barlaz M (2009) Accumulation and fragmentation of plastic debris in global environments. Philos Trans R Soc B 364:1985–1998

Barreiros JP, Raykov VS (2014) Lethal lesions and amputation caused by plastic debris and fishing gear on the loggerhead turtle *Caretta caretta* (Linnaeus, 1758). Three case reports from Terceira Island, Azores (NE Atlantic). Mar Pollut Bull 86:518–522

Barrios-Garrido H, Petit-Rodriguez MJ, Moreno E, Wildermann N (2013) Ghost nets: a new hazard to sea turtles in the Gulf of Venezuela. In: Tucker T, Belskis L, Panagopoulou A, Rees AL and others (eds) Proc 33rd Annu Symp Sea Turtle Biology and Conservation, 5–8 Feb 2013, Baltimore, MD. NOAA Tech Memo NMFS-SEFSC-645. Southeast Fisheries Science Center, Miami, FL, p 89

Bentivegna F (1995) Endoscopic removal of polyethylene cord from a loggerhead turtle. Mar Turtle Newsl 71:5

Blasi MF, Mattei D (2017) Seasonal encounter rate, life stages and main threats to the loggerhead sea turtle (*Caretta caretta*) in the Aeolian Archipelago (southern Thyrrenian Sea). Aquat Conserv 27:617–630

Bolten AB (2003) Variation in sea turtle life history patterns: neritic vs. oceanic developmental stages. In: Lutz PL, Musick JA, Wyneken J (eds) The biology of sea turtles, Vol 2. CRC Press, Boca Raton, FL, p 243–257

Burgman M, Carr A, Godden L, Gregory R, McBride M, Flander L, Maguire L (2011a) Redefining expertise and improving ecological judgment. Conserv Lett 4:81–87

Burgman MA, McBride M, Ashton R, Speirs-Bridge A and others (2011b) Expert status and performance. PLOS ONE 6:e22998

Camedda A, Marra S, Matiddi M, Massaro G and others (2014) Interaction between loggerhead sea turtles (*Caretta caretta*) and marine litter in Sardinia (Western Mediterranean Sea). Mar Environ Res 100:25–32

Casale P (2011) Sea turtle by-catch in the Mediterranean. Fish Fish 12:299–316

Casale P, Affronte M, Insacco G, Freggi D and others (2010) Sea turtle strandings reveal high anthropogenic mortality in Italian waters. Aquat Conserv 20:611–620

Casale P, Freggi D, Paduano V, Oliverio M (2016) Biases and best approaches for assessing debris ingestion in sea turtles, with a case study in the Mediterranean. Mar Pollut Bull 110:238–249

Chaloupka M, Work TM, Balazs GH, Murakawa SKK, Morris R(2008) Cause-specific temporal and spatial trends in green sea turtle strandings in the Hawaiian Archipelago (1982–2003). Mar Biol 154:887–898

Chatto R (1995) Sea turtles killed by flotsam in northern Australia. Mar Turtle Newsl 69:17–18

Davies RWD, Cripps SJ, Nickson A, Porter G (2009) Defining and estimating global marine fisheries bycatch. Mar Policy 33:661–672

Elaine AI, Seaman CA (2007) Likert scales and data analyses. Qual Prog 40:64–65

Elo S, Kyngäs H (2008) The qualitative content analysis process. J Adv Nurs 62:107–115

Epperly SP, Braun J, Chester AJ, Cross FA, Merriner JV, Tester PA, Churchill JH (1996) Beach strandings as an indicator of at-sea mortality of sea turtles. Bull Mar Sci 59:289–297

Filmalter JD, Capello M, Deneubourg JL, Cowley PD, Dagorn L (2013) Looking behind the curtain: quantifying massive shark mortality in fish aggregating devices. Front Ecol Environ 11:291–296

Finkbeiner EM, Wallace BP, Moore JE, Lewison RL, Crowder LB, Read AJ (2011) Cumulative estimates of sea turtle bycatch and mortality in USA fisheries between 1990 and 2007. Biol Conserv 144:2719–2727

Francke DL, Balazs GH, Brunson S, Nurzia Humburg I and others (2014) Marine turtle strandings in the Hawaiian Islands January-December 2013. NOAA Pacific Islands Fisheries Science Centre Internal Report IR-14-003. PIFSC, Honolulu, HI

Gall SC, Thompson RC (2015) The impact of debris on marine life. Mar Pollut Bull 92:170–179

Gardner B, Sullivan PJ, Morreale SJ, Epperly SP (2008) Spatial and temporal statistical analysis of bycatch data: patterns of sea turtle bycatch in the North Atlantic. Can J Fish Aquat Sci 65:2461–2470

Gilman EL (2011) Bycatch governance and best practice mitigation technology in global tuna fisheries. Mar Policy 35:590–609

Gilman E, Suuronen P, Hall M, Kennelly S (2013) Causes and methods to estimate cryptic sources of fishing mortality. J Fish Biol 83:766–803

Gilman E, Chopin F, Suuronen P, Kuemlangan B (2016) Abandoned, lost and discarded gillnets and trammel nets. Methods to estimate ghost fishing mortality, and status of regional monitoring and management. FAO Fisheries and Aquaculture Technical Paper No. 600. FAO, Rome

Good TP, June JA, Etnier MA, Broadhurst G (2010) Derelict fishing nets in Puget Sound and the Northwest Straits: patterns and threats to marine fauna. Mar Pollut Bull 60: 39–50

Hamann M, Godfrey M, Seminoff J, Arthur K and others (2010) Global research priorities for sea turtles: informing management and conservation in the 21st century. Endang Species Res 11:245–269

Hamann M, Grech A, Wolanski E, Lambrechts J (2011) Modelling the fate of marine turtle hatchlings. Ecol Modell 222:1515–1521

Hart KM, Mooreside P, Crowder L (2006) Interpreting the spatio-temporal patterns of sea turtle strandings: going with the flow. Biol Conserv 129:283–290

Hunt KE, Innis CJ, Merigo C, Rolland RM (2016) Endocrine responses to diverse stressors of capture, entanglement and stranding in leatherback turtles (*Dermochelys coriacea*). Conserv Physiol 4:cow022

Innis C, Merigo C, Dodge K, Tlusty M and others (2010) Health evaluation of leatherback turtles (*Dermochelys*

coriacea) in the northwestern Atlantic during direct capture and fisheries gear disentanglement. Chelonian Conserv Biol 9:205–222

Jambeck JR, Geyer R, Wilcox C, Siegler TR and others (2015) Plastic waste inputs from land into the ocean. Science 347:768–771

Jeffers VF, Godley BJ (2016) Satellite tracking in sea turtles: How do we find our way to the conservation dividends? Biol Conserv 199:172–184

Jensen MP, Abreu-Grobois FA, Frydenberg J, Loeschcke V (2006) Microsatellites provide insight into contrasting mating patterns in arribada vs. non-arribada olive ridley sea turtle rookeries. Mol Ecol 15:2567–2575

Jensen M, Limpus C, Whiting S, Guinea M and others (2013) Defining olive ridley turtle *Lepidochelys olivacea* management units in Australia and assessing the potential impact of mortality in ghost nets. Endang Species Res 21: 241–253

Koch V, Nichols WJ, Peckham H, de la Toba V (2006) Estimates of sea turtle mortality from poaching and bycatch in Bahía Magdalena, Baja California Sur, Mexico. Biol Conserv 128:327–334

Laist DW (1987) Overview of the biological effects of lost and discarded plastic debris in the marine environment. Mar Pollut Bull 18:319–326

Lawson TJ, Wilcox C, Johns K, Dann P, Hardesty BD (2015) Characteristics of marine debris that entangle Australian fur seals (*Arctocephalus pusillus doriferus*) in southern Australia. Mar Pollut Bull 98:354–357

López-Jurado LF, Varo-Cruz N, Lopez-Suarez P (2003) Incidental capture of loggerhead turtles (*Caretta caretta*) on Boa Vista (Cape Verde Islands). Mar Turtle Newsl 101: 14–16

Macfadyen G, Huntington T, Cappell R (2009) Abandoned, lost or otherwise discarded fishing gear. UNEP Regional Seas Reports and Studies No. 185, FAO Tech Pap No. 523. FAO, Rome

Martin TG, Burgman MA, Fidler F, Kuhnert PM , Low-Choy S, McBride M, Mengersen K (2012) Eliciting expert knowledge in conservation science. Conserv Biol 26: 29–38

Matsuoka T, Nakashima T, Nagasawa N (2005) A review of ghost fishing: scientific approaches to evaluation and solutions. Fish Sci 71:691–702

Mazaris AD, Fiksen Ø, Matsinos YG (2005) Using an individual-based model for assessment of sea turtle population viability. Popul Ecol 47:179–191

McIntosh RR, Kirkwood R, Sutherland DR, Dann P (2015) Drivers and annual estimates of marine wildlife entanglement rates: a long-term case study with Australian fur seals. Mar Pollut Bull 101:716–725

McMahon C, Bradshaw C, Hays G (2007) Satellite tracking reveals unusual diving characteristics for a marine reptile, the olive ridley turtle *Lepidochelys olivacea*. Mar Ecol Prog Ser 329:239–252

Meager JJ, Limpus CJ (2012) Marine wildlife stranding and mortality database annual report 2011. III. Marine turtle. Conservation Technical and Data Report 2012. Department of Environment and Heritage Protection, Brisbane

Moore E, Lyday S, Roletto J, Litle K and others (2009) Entanglements of marine mammal and sea birds in central California and the north-west coast of the United States 2001–2005. Mar Pollut Bull 58:1045–1051

MSFD GES Technical Subgroup on Marine Litter (2011) Marine litter: technical recommendations for the implementation of MSFD requirements. European Commission Joint Research Centre and Institute for Environment and Sustainability, Luxembourg doi:10.2788/92438

Nelms SE, Duncan EM, Broderick AC, Galloway TS and others (2016) Plastic and marine turtles: a review and call for research. ICES J Mar Sci 73:165–181

Newing H (2011) Conducting research in conservation: social science methods and practice. Routledge, New York, NY

Orós J, Montesdeoca N, Camacho M, Arencibia A, Calabuig P (2016) Causes of stranding and mortality, and final disposition of loggerhead sea turtles (*Caretta caretta*) admitted to a wildlife rehabilitation center in Gran Canaria Island, Spain (1998-2014): a long-term retrospective study. PLOS ONE 11:e0149398

Polovina JJ, Balazs GH, Howell EA, Parker DM, Seki MP, Dutton PH (2004) Forage and migration habitat of loggerhead (*Caretta caretta*) and olive ridley (*Lepidochelys olivacea*) sea turtles in the central North Pacific Ocean. Fish Oceanogr 13:36–51

Raum-Suryan KL, Jemison LA, Pitcher KW (2009) Entanglement of Steller sea lions (*Eumetopias jubatus*) in marine debris: identifying causes and finding solutions. Mar Pollut Bull 58:1487–1495

Rees AF, Al Saady S, Broderick AC, Coyne MS, Papathanasopoulou N, Godley BJ (2010) Behavioural polymorphism in one of the world's largest populations of loggerhead sea turtles *Caretta caretta*. Mar Ecol Prog Ser 418: 201–212

Rees AF, Alfaro-Shigueto J, Barata PCR, Bjorndal KA and others (2016) Are we working towards global research priorities for management and conservation of sea turtles? Endang Species Res 31:337–382

Remie S, Mortimer JA (2007) First records of olive ridley turtles (*Lepidochelys olivacea*) in Seychelles. Mar Turtle Newsl 117:9

Russo G, Di Bella C, Loria GR, Insacco G, Palazzo P, Violani C, Zava B (2014) Notes on the influence of human activities on sea chelonians in Sicilian waters. J Mt Ecol 7(Suppl):37–41

Ryan PG, Cole G, Spiby K, Nel R, Osborne A, Perold V (2016) Impacts of plastic ingestion on post-hatchling loggerhead turtles off South Africa. Mar Pollut Bull 107: 155–160

Saldanha HJ, Sancho G, Santos MN, Puente E and others (2003) The use of biofouling for ageing lost nets: a case study. Fish Res 64:141–150

Santos AJB, Bellini C, Bortolon LF, Coluchi R (2012) Ghost nets haunt the olive ridley turtle (*Lepidochelys olivacea*) near the Brazilian Islands of Fernando de Noronha and Atol das Rocas. Herpetol Rev 43:245–246

Santos RG, Andrades R, Boldrini MA, Martins AS (2015) Debris ingestion by juvenile marine turtles: an underestimated problem. Mar Pollut Bull 93:37–43

Santos RG, Andrades R, Fardim LM, Martins AS (2016) Marine debris ingestion and Thayer's law — the importance of plastic color. Environ Pollut 214:585–588

Sasso CR, Epperly SP (2007) Survival of pelagic juvenile loggerhead turtles in the open ocean. J Wildl Manag 71: 1830–1835

Schuyler Q, Hardesty BD, Wilcox C, Townsend K (2014) Global analysis of anthropogenic debris ingestion by sea turtles. Conserv Biol 28:129–139

Smolowitz RJ (1978) Lobster, *Homarus americanus*, trap design and ghost fishing. Mar Fish Rev 40:59–67

Stelfox M, Hudgins J (2015) A two year summary of turtle entanglements in ghost gear in the Maldives. Indian Ocean Turtle Newsletter 22:14–20

Stelfox M, Hudgins J, Sweet M (2016) A review of ghost gear entanglement amongst marine mammals, reptiles and elasmobranchs. Mar Pollut Bull 111:6–17

Tasker M, Camphuysen CJ, Cooper J, Garthe S, Montevecchi WA, Blaber SJM (2000) The impacts of fishing on marine birds. ICES J Mar Sci 57:531–547

Thompson RC, Olsen Y, Mitchell RP, Davis A and others (2004) Lost at sea: Where is all the plastic? Science 304: 838

Vegter A, Barletta M, Beck C, Borrero J and others (2014) Global research priorities to mitigate plastic pollution impacts on marine wildlife. Endang Species Res 25: 225–247

Votier SC, Archibald K, Morgan G, Morgan L (2011) The use of plastic debris as nesting material by a colonial seabird and associated entanglement mortality. Mar Pollut Bull 62:168–172

Wallace BP, DiMatteo AD, Hurley BJ, Finkbeiner EM and others (2010a) Regional management units for marine turtles: a novel framework for prioritizing conservation and research across multiple scales. PLOS ONE 5:e15465

Wallace BP, Lewison RL, McDonald SL, McDonald RK and others (2010b) Global patterns of marine turtle bycatch. Conserv Lett 3:131–142

White D (2006) Marine debris in Northern Territory waters 2004. WWF Australia, Sydney

Wilcox C, Hardesty BD, Sharples R, Griffin DA, Lawson TJ, Gunn R (2013) Ghostnet impacts on globally threatened turtles, a spatial risk analysis for northern Australia. Conserv Lett 6:247–254

Wilcox C, Heathcote G, Goldberg J, Gunn R, Peel D, Hardesty BD (2015) Understanding the sources and effects of abandoned, lost, and discarded fishing gear on marine turtles in northern Australia. Conserv Biol 29:198–206

Wong SL, Ngadi N, Abdullah TAT, Inuwa IM (2015) Current state and future prospects of plastic waste as source of fuel: a review. Renew Sustain Energy Rev 50:1167–1180

Editorial responsibility: Rory Wilson, Swansea, UK

Submitted: February 22, 2017; Accepted: September 22, 2017
Proofs received from author(s): November 28, 2017

De Sitter entropy as holographic entanglement entropy

Nikolaos Tetradis[⋆]

Department of Physics, National and Kapodistrian University of Athens,
University Campus, Zographou 157 84, Greece

⋆ ntetrad@phys.uoa.gr

4th International Conference on Holography,
String Theory and Discrete Approach
Hanoi, Vietnam, 2020
doi:10.21468/SciPostPhysProc.4

Abstract

We review the results of refs. [1, 2], in which the entanglement entropy in spaces with horizons, such as Rindler or de Sitter space, is computed using holography. This is achieved through an appropriate slicing of anti-de Sitter space and the implementation of a UV cutoff. When the entangling surface coincides with the horizon of the boundary metric, the entanglement entropy can be identified with the standard gravitational entropy of the space. For this to hold, the effective Newton's constant must be defined appropriately by absorbing the UV cutoff. Conversely, the UV cutoff can be expressed in terms of the effective Planck mass and the number of degrees of freedom of the dual theory. For de Sitter space, the entropy is equal to the Wald entropy for an effective action that includes the higher-curvature terms associated with the conformal anomaly. The entanglement entropy takes the expected form of the de Sitter entropy, including logarithmic corrections.

Received 22-10-2020
Accepted 09-11-2020
Published 13-08-2021
doi:10.21468/SciPostPhysProc.4.002

The fact that the divergent part of the entanglement entropy scales with the area of the entangling surface [3, 4] suggests a connection with the gravitational entropy of spaces containing horizons. It seems reasonable that the entropies should become equal when the entangling surface is identified with a horizon. We address this problem in the context of the AdS/CFT correspondence through use of appropriate coordinates that set the boundary metric in Rindler or static de Sitter form. According to the Ryu-Takayanagi proposal [5–7], the entanglement entropy of a part of the AdS boundary within an entangling surface $\mathcal{A}$ is proportional to the area of a minimal surface γ_A anchored on $\mathcal{A}$ and extending into the bulk.

We consider the standard parameterization of $(d+2)$-dimensional AdS space with global coordinates, as well as parametrizations through Fefferman-Graham coordinates, with the boundary located at the value $z = 0$ of the bulk coordinate. As a first case we consider a metric with a Rindler boundary:

$$ds^2_{d+2} = \frac{R^2}{z^2}\left[dz^2 - a^2 y^2 d\eta^2 + dy^2 + d\vec{x}_{d-1}\right], \tag{1}$$

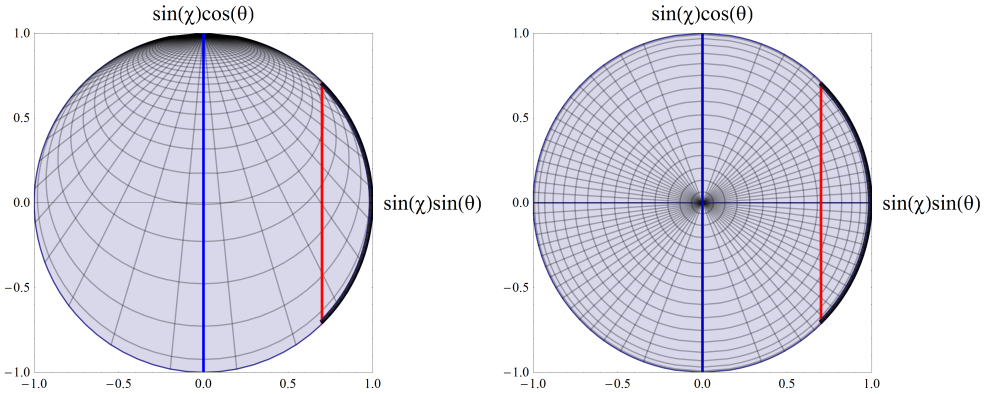


Figure 1: Constant-time slice of AdS_3 for a Rindler boundary with $a = 1$ (left) and a static de Sitter boundary with $H = 1$ (right).

where a is a constant parameter. The timelike coordinate η takes values $-\infty < \eta < \infty$. The range $0 < y < \infty$ of the spacelike coordinate y covers the right (R) Rindler wedge, while the range $-\infty < y < 0$ covers the left (L) wedge.

In the left plot of fig. 1 we depict how the slice of the AdS_3 cylinder with $\eta = 0$ is covered by the coordinates y and z for $a = 1$. The two axes correspond to global coordinates. The circumference is the AdS_3 boundary with $z = 0$, which is parameterized by the coordinate y. The Rindler horizon at $y = 0$ corresponds to the point $(0,-1)$ in fig. 1. Positive values of y cover the right semicircle (R wedge), and negative values the left semicircle (L wedge). The point $(0,1)$ is approached in the limits $y \to \pm\infty$ from right or left. The AdS_3 interior is covered by lines of constant y and variable positive z. All these lines converge to the point (0,1) for $z \to \infty$. We expect to have entanglement between the R and L wedges. The corresponding entanglement entropy can obtained through holography by computing the area of the minimal surface γ_A of ref. [5–7]. This is depicted by the blue line in this case, which acts as a bulk horizon. The Rindler horizon can be viewed as the holographic image of the bulk horizon.

Let us consider a strip with width l in the y-direction and very large extent in the remaining spacelike directions. The minimal surface extends into the bulk up to $z_* = \Gamma\left(\frac{1}{2d}\right)/\left(2\sqrt{\pi}\,\Gamma\left(\frac{d+1}{2d}\right)\right) l$. In global coordinates this surface corresponds to a straight line through the bulk, as depicted by the red line in fig. 1. The entanglement entropy can be computed as

$$S_A = \frac{2R(R^{d-1}L^{d-1})}{4G_{d+2}}\left(\frac{1}{(d-1)\epsilon^{d-1}} + \frac{\sqrt{\pi}}{2d}\frac{\Gamma\left(\frac{1-d}{2d}\right)}{\Gamma\left(\frac{1}{2d}\right)}\frac{1}{z_*^{d-1}}\right). \tag{2}$$

A cutoff ϵ has been imposed on z as the surface approaches the boundary. For $d = 1$, one must substitute $1/((d-1)\epsilon^{d-1})$ with $\log(1/\epsilon)$. Here L is the large length of the directions perpendicular to the strip, so that $R^{d-1}L^{d-1}$ is the corresponding volume.

We are interested in the limit in which the width l of the strip covers the whole R wedge. In this case the entanglement occurs between the R and L wedges. For $l \to \infty$ we have $z_* \to \infty$ and the second term in the parenthesis in eq. (2) vanishes. In order to assign a physical meaning to the first term, we can define the effective Newton's constant for the boundary theory as in [8]:

$$G_{d+1} = (d-1)\epsilon^{d-1}\frac{G_{d+2}}{R}, \tag{3}$$

with $(d-1)\epsilon^{d-1}$ replaced by $1/\log(1/\epsilon)$ for $d = 1$. This definition can be justified in the context of the Randall-Sundrum (RS) model [9, 10], which employs only the part of the AdS

space with $z > \epsilon$. The effective low-energy theory includes dynamical gravity with a Newton's constant given by eq. (3). In the limit $\epsilon \to 0$, the constant vanishes and gravity becomes non-dynamical. This demonstrates the difficulty in computing the gravitational entropy in the context of the AdS/CFT coorespondence. The resolution we suggest is to keep the cutoff nonzero and absorb it in the definition of the effective Newton's constant. Trading ϵ for G_{d+1} in the expression for the entropy results in a meaningful expression.

Substituting eq. (3) in eq. (2) gives an entanglement entropy which is bigger by a factor of 2 than the known gravitational entropy [11]. The reason can be traced to the way the limit is taken in order to cover the whole R wedge. We start from a strip in the y-direction extending between two points y_1 and y_2, and then take the limits $y_1 \to 0$ and $y_2 \to \infty$. The first limit leads to the location of the Rindler horizon. However, any finite value of y_2 excludes an infinite domain corresponding to $y > y_2$. As a result, the strip is entangled not only with the (infinite) L wedge, but also with the (infinite) domain $y > y_2$. The two contributions are expected to be equal because the space is essentially flat. Obtaining the entropy corresponding to the entanglement with the L wedge only can be obtained by dividing the result with a factor of 2. The final result for the Rindler entropy is

$$S_R = \frac{R^{d-1} L^{d-1}}{4 G_{d+1}}, \tag{4}$$

in agreement with [11]. It is also illuminating to observe that the bulk horizon depicted as a blue line in fig. 1 approaches the boundary at two points. The point $(0,-1)$ is the true Rindler horizon. However, the point $(0,1)$ does not belong to the boundary Rindler space, but corresponds only to the limits $y \to \pm\infty$. The contribution to the area of the entangling surface from its vicinity should not be taken into account, thus justifying the division by 2.

The second case we consider is that of a boundary static de Sitter (dS) space:

$$ds^2_{d+2} = \frac{R^2}{z^2}\left[dz^2 + \left(1 - \frac{1}{4}H^2 z^2\right)^2 \left(-(1-H^2\rho^2)dt^2 + \frac{d\rho^2}{1-H^2\rho^2} + \rho^2\, d\Omega^2_{d-1}\right)\right]. \tag{5}$$

For $d > 1$, the range $0 \le \rho \le 1/H$ covers one static patch. There are two such patches in the global geometry, which start from the the "North" or "South pole" at $\rho = 0$ and are joined at the surface with $\rho = 1/H$. For $d = 1$, ρ can also take negative value and each static patch is covered by $-1/H \le \rho \le 1/H$. In the right plot of fig. 1 we depict how the slice of the AdS_3 cylinder with $t = 0$ is covered by the coordinates ρ and z for $H = 1$. The circumference is again the AdS_3 boundary with $z = 0$, which is parameterized by the coordinate ρ. There are two horizons: one at $\rho = -1$, corresponding to the point $(0,-1)$, and one at $\rho = 1$, corresponding to the point $(0,1)$ on the boundary. The AdS_3 interior is covered by lines of constant ρ and variable positive z. All these lines converge to the point (0,0) at the center for $z \to \infty$. In the context of the global geometry, we expect to have entanglement between the two static patches. The corresponding entanglement entropy can be obtained through holography by computing the area of the minimal surface γ_A of ref. [5–7], depicted by the blue line. This line acts as bulk horizon. The difference with the Rindler case we discussed before is that the endpoints of the minimal surface are points of the boundary dS space, they are actually the horizons. This means that there is no need to divide by a factor of 2 in this case. For $d > 1$ the d-dimensional minimal surface γ_A ends on an $(d-1)$-dimensional sphere that separates the two hemispheres of the slice of dS_{d+1} with $t = 0$.

The isometries of dS space indicate that the entangling surface is spherical in this case. The minimal surface γ_A in the bulk can be determined by minimizing the integral

$$\text{Area}(\gamma_A) = R^d S^{d-1} \int d\sigma \frac{\sin^{d-1}(\sigma)}{\sinh^d(w)} \sqrt{1 + \left(\frac{dw(\sigma)}{d\sigma}\right)^2}, \tag{6}$$

where we have defined the parameters $\sigma = \sin^{-1}(H\rho)$, $w = 2\tanh^{-1}(Hz/2)$, and denoted the volume of the $(d-1)$-dimensional unit sphere as S^{d-1}. The above expression is minimized by the function [2]

$$w(\sigma) = \cosh^{-1}\left(\frac{\cos(\sigma)}{\cos(\sigma_0)}\right). \tag{7}$$

For $\sigma_0 \to 0$ the known expression $w(\sigma) = \sqrt{\sigma_0^2 - \sigma^2}$ [5–7] for $H = 0$ is reproduced. For $\sigma_0 \to \pi/2$ the boundary is approached at the location of the horizon with $dw/d\sigma \to -\infty$.

The integral (6) is dominated by the region near the boundary. Introducing a cutoff at $z = \epsilon$ results in a leading contribution

$$\text{Area}(\gamma_A) = R^d S^{d-1} I(\epsilon) = R^d S^{d-1} \int_{H\epsilon} \frac{dw}{\sinh^d(w)}. \tag{8}$$

For $d \neq 1$ the leading divergent part is $I(\epsilon) = 1/((d-1)H^{d-1}\epsilon^{d-1})$, while for $d = 1$ it is $\log(1/(H\epsilon))$. Using eq. (3) we obtain the leading contribution to the entropy:

$$S_{\text{dS}} = \frac{\text{Area}(\gamma_A)}{4G_{d+2}} = \frac{R^d S^{d-1}}{4G_{d+2}(d-1)H^{d-1}\epsilon^{d-1}} = \frac{S^{d-1}}{4G_{d+1}}\left(\frac{R}{H}\right)^{d-1} = \frac{A_H}{4G_{d+1}}, \tag{9}$$

with A_H the area of the horizon. This result reproduces the gravitational entropy of [12]. It is valid for $d = 1$ as well, with $1/((d-1)\epsilon^{d-1})$ replaced by $\log(1/\epsilon)$ and $S^0 = 2$, because the horizons of the global dS_2 geometry are 2 points [8].

The integral $I(\epsilon)$ also contains subleading divergences. There is a subleading logarithmic divergence for $d = 3$, no singular subleading terms for $d = 2$, while the only divergence for $d = 1$ is the leading logarithmic term already included in eq. (9). For $d > 3$ we have subleading power-law divergences for odd $d+1$, plus a logarithmic one for even $d+1$. We focus on four dimensions, in which the dS entropy takes the form

$$S_{\text{dS}} = \frac{A_H}{4G_4}\left(1 + H^2\epsilon^2 \log H\epsilon\right). \tag{10}$$

The logarithmic dependence on the cutoff hints at a connection with the conformal anomaly of the dual theory, which results from higher curvature terms in the effective theory. The effective action can be deduced from known results for the on-shell action in holographic renormalization [13–15]. In our approach the divergences are not removed through the introduction of counterterms, but are absorbed in the effective couplings. This means that the relevant quantity for our purposes is the regulated form of the effective action. Using the results of [13–15], we obtain the leading terms [2]

$$S = \frac{R^3}{16\pi G_5}\int d^4x\,\sqrt{-\gamma}\left[\frac{6}{\epsilon^4} + \frac{1}{2\epsilon^2}\mathcal{R} - \frac{1}{4}\log\epsilon\left(\mathcal{R}_{ij}\mathcal{R}^{ij} - \frac{1}{3}\mathcal{R}^2\right)\right]. \tag{11}$$

The first term corresponds to a cosmological constant. In the RS model [9,10] this is balanced by the surface tension of the brane at $z = \epsilon$. The second term is the standard Einstein term if the effective Newton's constant G_4 is defined as in eq. (3) with $d = 3$. The third term is responsible for the holographic conformal anomaly. The action (11) supports a dS solution. In order to take into account the presence of the higher-curvature terms in eq. (11) one must compute the Wald entropy [16–18]. The result is in agreement with the singular part of the correction provided by the holographic calculation (10) [2].

For the $\mathcal{N} = 4$ supersymmetric $SU(N)$ gauge theory in the large-N limit, the effective action can be computed as [19]

$$S = -\frac{\beta}{16\pi^2}\Gamma\left(2 - \frac{d+1}{2}\right)\int d^4x\,\sqrt{-\gamma}\left(\mathcal{R}_{ij}\mathcal{R}^{ij} - \frac{1}{3}\mathcal{R}^2\right), \tag{12}$$

with $\beta = -N^2/4$. The divergence of $\Gamma(2-(d+1)/2)$ in dimensional regularization in the limit $d+1 \to 4$ corresponds to a $\log(1/\epsilon^2)$ divergence in our cutoff regularization. A comparison of the above expression with eq. (11) reproduces the standard AdS/CFT relation $G_5 = \pi R^3/(2N^2)$. The dimensionful UV momentum cutoff for $d = 3$ can be expressed as $(\epsilon_N R)^{-2} = 2G_5/(R^3 G_4) = 8\pi^2 m_{\rm Pl}^2/N^2$, with $m_{\rm Pl}^2 = 1/(8\pi G_4)$. Now eq. (10) for $d = 3$ can be cast in the form

$$S_{\rm dS} = \frac{A_H}{4G_4} + N^2 \log(H\epsilon_N) = \frac{A_H}{4G_4} + N^2 \log\left(\frac{N}{\sqrt{8}\pi}\frac{H/R}{m_{\rm Pl}}\right), \tag{13}$$

where H/R is the physical Hubble scale. This expression is completely analogous to the black-hole result [20], with the horizon size parameter measured in units of the UV cutoff. It is also in agreement with the calculation of the logarithmic part of the holographic entanglement entropy in [21].

The calculation of the entropy associated with nontrivial gravitational backgrounds through holography faces two difficulties:

- The boundary metric in the context of AdS/CFT is not dynamical, a feature that is equivalent to $m_{\rm Pl} \to \infty$.
- The entanglement entropy has a strong dependence on the UV cutoff of the theory, which makes its identification with the gravitational entropy problematic.

We showed that these difficulties can be resolved if the UV cutoff dependence is absorbed in the definition of $m_{\rm Pl}$. The conceptual framework is provided by the Randall-Sundrum model [9, 10], or, alternatively, by the regulated form of the effective action in holographic renormalization [13–15]. Our derivation of the dS entropy is consistent with the expectation that the entropy associated with gravitational horizons can be understood as entanglement entropy if Newton's constant is induced by quantum fluctuations of matter fields [22, 23]. In the context of the AdS/CFT correspondence the bulk degrees of freedom correspond to the matter fields of the dual theory. The boundary Einstein action arises through the integration of these bulk degrees of freedom up to the UV cutoff.

Our approach is in contrast with the usual interpretation of the leading contribution to the entanglement entropy as an unphysical UV-dependent quantity of little interest. We have reached the opposite conclusion: The leading contribution to the entropy has a universal form that depends only on the horizon area because the same degrees of freedom contribute to the entropy and Newton's constant. Also, the detailed nature of the UV cutoff does not affect the leading contribution. The particular features of the underlying theory, such as the number of degrees of freedom become apparent at the level of the subleading corrections to the entropy: the coefficient of the logarithmic correction is determined by the central charge of the theory.

Acknowledgments

This research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the First Call for H.F.R.I. Research Projects to support Faculty members and Researchers (Project Number: 824).

References

[1] D. Giataganas and N. Tetradis, *Entanglement entropy, horizons and holography*, Phys. Lett. B **796**, 88 (2019), doi:10.1016/j.physletb.2019.07.019.

[2] N. Tetradis, *Corrections to de Sitter entropy through holography*, Phys. Lett. B **807**, 135552 (2020), doi:10.1016/j.physletb.2020.135552.

[3] L. Bombelli, R. K. Koul, J. Lee and R. D. Sorkin, *Quantum source of entropy for black holes*, Phys. Rev. D **34**, 373 (1986), doi:10.1103/PhysRevD.34.373.

[4] M. Srednicki, *Entropy and area*, Phys. Rev. Lett. **71**, 666 (1993), doi:10.1103/PhysRevLett.71.666.

[5] S. Ryu and T. Takayanagi, *Holographic derivation of entanglement entropy from AdS/CFT*, Phys. Rev. Lett. **96**, 181602 (2006), doi:10.1103/PhysRevLett.96.181602.

[6] T. Nishioka, S. Ryu and T. Takayanagi, *Holographic entanglement entropy: An overview*, J. Phys. A **42**, 504008 (2009), doi:10.1088/1751-8113/42/50/504008.

[7] S. Ryu and T. Takayanagi, *Aspects of holographic entanglement entropy*, J. High Energy Phys. **08**, 045 (2006), doi:10.1088/1126-6708/2006/08/045.

[8] S. Hawking, J. M. Maldacena and A. Strominger, *de Sitter entropy, quantum entanglement and AdS/CFT*, J. High Energy Phys. **05**, 001 (2001), doi:10.1088/1126-6708/2001/05/001.

[9] L. Randall and R. Sundrum, *A large mass hierarchy from a small extra dimension*, Phys. Rev. Lett. **83**, 3370 (1999), doi:10.1103/PhysRevLett.83.3370.

[10] L. Randall and R. Sundrum, *An alternative to compactification*, Phys. Rev. Lett. **83**, 4690 (1999), doi:10.1103/PhysRevLett.83.4690.

[11] R. Laflamme, *Entropy of a Rindler wedge*, Phys. Lett. B **196**, 449 (1987), doi:10.1016/0370-2693(87)90799-4.

[12] G. W. Gibbons and S. W. Hawking, *Cosmological event horizons, thermodynamics, and particle creation*, Phys. Rev. D **15**, 2738 (1977), doi:10.1103/PhysRevD.15.2738.

[13] S. de Haro, K. Skenderis and S. N. Solodukhin, *Holographic reconstruction of spacetime and renormalization in the AdS/CFT correspondence*, Commun. Math. Phys. **217**, 595 (2001), doi:10.1007/s002200100381.

[14] K. Skenderis, *Lecture notes on holographic renormalization*, Class. Quant. Grav. **19**, 5849 (2002), doi:10.1088/0264-9381/19/22/306.

[15] I. Papadimitriou and K. Skenderis, *AdS/CFT correspondence and geometry*, IRMA Lect. Math. Theor. Phys. **8**, 73 (2005), doi:10.4171/013-1/4.

[16] R. M. Wald, *Black hole entropy is the Noether charge*, Phys. Rev. D **48**, R3427 (1993), doi:10.1103/PhysRevD.48.R3427.

[17] V. Iyer and R. M. Wald, *Some properties of the Noether charge and a proposal for dynamical black hole entropy*, Phys. Rev. D **50**, 846 (1994), doi:10.1103/PhysRevD.50.846.

[18] T. Jacobson, G. Kang and R. C. Myers, *On black hole entropy*, Phys. Rev. D **49**, 6587 (1994), doi:10.1103/PhysRevD.49.6587.

[19] N. D. Birrell and P. C. W. Davies, *Quantum fields in curved space*, Cambridge Monographs on Mathematical Physics, Cambridge University Press (1982).

[20] A. Sen, *Logarithmic corrections to Schwarzschild and other non-extremal black hole entropy in different dimensions*, J. High Energ. Phys. **04**, 156 (2013), doi:10.1007/JHEP04(2013)156.

[21] H. Casini, M. Huerta and R. C. Myers, *Towards a derivation of holographic entanglement entropy*, J. High Energ. Phys. **05**, 036 (2011), doi:10.1007/JHEP05(2011)036.

[22] L. Susskind and J. Uglum, *Black hole entropy in canonical quantum gravity and superstring theory*, Phys. Rev. D **50**, 2700 (1994), doi:10.1103/PhysRevD.50.2700.

[23] T. Jacobson, *Black hole entropy and induced gravity* (1994), arXiv:gr-qc/9404039.

Escalation through Entanglement

How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War

James M. Acton

The 2018 U.S. Nuclear Posture Review contains a highly consequential threat that has been largely overlooked in the wave of commentary surrounding the document's release: the United States warns potential adversaries that it would consider using nuclear weapons in the event of "significant nonnuclear strategic attacks . . . on U.S. or allied nuclear forces, their command and control, or warning and attack assessment capabilities."[1] This threat was motivated by the growing vulnerability of these assets—in particular, the United States' nuclear command, control, communication, and intelligence (C3I or enabling) capabilities—to advanced nonnuclear weapons, and is presumably intended to deter attacks on them.[2] In issuing this threat, the Nuclear Posture Review illustrates that nonnuclear attacks on nuclear forces and C3I capabilities could be highly escalatory, even to the point of directly sparking a nuclear war.

A key challenge in managing these escalation risks is that attacks on an opponent's nuclear forces or their C3I capabilities (whether they belong to the United States or another state) might not be deliberate. Since the late 2000s, scholars have warned about the possibility of escalation in a U.S.-China conflict resulting from so-called crisis instability generated by actual or threatened U.S. nonnuclear operations that were intended to suppress China's conventional forces but inadvertently degraded its nuclear forces or associated C3I assets located in the theater of operations, thus leading Beijing to fear it was

James M. Acton is co-director of the Nuclear Policy Program and holds the Jessica T. Matthews Chair at the Carnegie Endowment for International Peace.

For insightful comments on previous drafts of this article, the author thanks Alexey Arbatov, Toby Dalton, Catherine Dill, Geoffrey Forden, Michael Gerson, Charles Glaser, Ariel Levite, Jeffrey Lewis, Li Bin, Austin Long, Tim Maurer, James Miller, George Perkovich, Pavel Podvig, Joshua Pollack, Brad Roberts, Scott Sagan, Petr Topychkanov, Tong Zhao, and the anonymous reviewers, as well as interviewees and participants at seminars where he presented this research. He is also grateful to Jessica Margolis, William Ossoff, Thu-An Pham, Kathryn Taylor, Elizabeth Whitfield, and Lauryn Williams for research assistance. This work received generous financial support from the Carnegie Corporation of New York. The contents of this article are exclusively the author's responsibility.

1. U.S. Department of Defense, "Nuclear Posture Review" (Washington, D.C.: U.S. Department of Defense, February 21, 2018), p. 21, https://media.defense.gov/2018/Feb/02/2001872886/-1/-1/1/2018-NUCLEAR-POSTURE-REVIEW-FINAL-REPORT.PDF.
2. Ibid., p. 56.

International Security, Vol. 43, No. 1 (Summer 2018), pp. 56–99, doi:10.1162/ISEC_a_00320

being disarmed.[3] Similar, if more infrequent, scholarly warnings have been voiced about a U.S.-Russia conflict.[4] Such escalation would be inadvertent because it was the result of military operations or threats that were not intended to be escalatory.[5]

This article's thesis is that the risks of inadvertent escalation are even more serious than these warnings suggest and are likely to increase significantly in the future. Driving these risks is the possibility that Chinese, Russian, or U.S. C3I assets located outside—potentially far outside—theaters of operation could be attacked over the course of a conventional conflict. These assets include satellites used for early warning, communication, and intelligence, surveillance, and reconnaissance (ISR); ground-based radars and transmitters; and communication aircraft.[6] Such assets constitute key nodes in states' nu-

3. Michael S. Chase, Andrew S. Erickson, and Christopher Yeaw, "Chinese Theater and Strategic Missile Force Modernization and Its Implications for the United States," *Journal of Strategic Studies*, Vol. 32, No. 1 (February 2009), pp. 101–106, doi:10.1080/01402390802407434; Jeffrey G. Lewis, "Chinese Nuclear Posture and Force Modernization," *Nonproliferation Review*, Vol. 16, No. 2 (July 2009), pp. 205–206, doi:10.1080/10736700902969661; Joshua Pollack, "Emerging Strategic Dilemmas in U.S.-Chinese Relations," *Bulletin of the Atomic Scientists*, Vol. 65, No. 4 (July/August 2009), pp. 53–63, doi:10.2968/065004006; Thomas J. Christensen, "The Meaning of the Nuclear Evolution: China's Strategic Modernization and U.S.-China Security Relations," *Journal of Strategic Studies*, Vol. 35, No. 4 (August 2012), pp. 467–471, doi:10.1080/01402390.2012.714710; Fiona S. Cunningham and M. Taylor Fravel, "Assuring Assured Retaliation: China's Nuclear Posture and U.S.-China Strategic Stability," *International Security*, Vol. 40, No. 2 (Fall 2015), pp. 40–45, doi:10.1162/ISEC_a_00215; Joshua H. Pollack, "Boost-Glide Weapons and U.S.-China Strategic Stability," *Nonproliferation Review*, Vol. 22, No. 2 (2015), pp. 157–161, doi:10.1080/10736700.2015.1119422; Wu Riqiang, "Sino-U.S. Inadvertent Escalation" (Atlanta: Program on Strategic Stability Evaluation, Georgia Institute of Technology, n.d.), https://www.yumpu.com/en/document/view/38495325/wu-sino-us-inadvertent-escalation-program-on-strategic-stability-; Caitlin Talmadge, "Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States," *International Security*, Vol. 41, No. 4 (Spring 2017), pp. 50–92, doi:10.1162/ISEC_a_00274; and Tong Zhao and Li Bin, "The Underappreciated Risks of Entanglement: A Chinese Perspective," in James M. Acton, ed., "Entanglement: Russian and Chinese Perspectives on Non-Nuclear Weapons and Nuclear Risks" (Washington, D.C.: Carnegie Endowment for International Peace, 2017), pp. 47–75, http://carnegieendowment.org/files/Entanglement_interior_FNL.pdf. A much larger literature, with many contributions from foreign authors, analyzes nonnuclear threats to nuclear forces but does not connect them to inadvertent escalation.

4. James M. Acton, "Silver Bullet? Asking the Right Questions about Conventional Prompt Global Strike" (Washington, D.C.: Carnegie Endowment for International Peace, 2013), pp. 120–126, http://carnegieendowment.org/files/cpgs.pdf; and Alexey Arbatov, Vladimir Dvorkin, and Petr Topychkanov, "Entanglement as a New Security Threat: A Russian Perspective," in Acton, *Entanglement*, pp. 9–45.

5. Forrest E. Morgan et al., *Dangerous Thresholds: Managing Escalation in the 21st Century* (Santa Monica, Calif.: RAND Corporation, 2008), pp. 23–25, http://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_MG614.pdf. This concept was first developed at length in Barry R. Posen, *Inadvertent Escalation: Conventional War and Nuclear Risks* (Ithaca, N.Y.: Cornell University Press, 1991), pp. 12–16.

6. Although threats to some of these assets have been discussed, their potential to spark inadvertent escalation has not.

clear C3I systems, but they are also "entangled" with nonnuclear weapons in two ways.[7] First, they are typically dual use; that is, they enable both nuclear and nonnuclear operations. Second, they are increasingly vulnerable to nonnuclear attack—much more vulnerable, in fact, than most nuclear-weapon delivery systems.

Entanglement could lead to escalation because both sides in a U.S.-Chinese or U.S.-Russian conflict could have strong incentives to attack the adversary's dual-use C3I capabilities to undermine its nonnuclear operations.[8] As a result, over the course of a conventional war, the nuclear C3I systems of one or both of the belligerents could become severely degraded. It is, therefore, not just U.S. nonnuclear strikes against China or Russia that could prove escalatory; Chinese or Russian strikes against American C3I assets could also—a possibility that scholars have scarcely even considered since the end of the Cold War.[9]

Two escalation mechanisms that have not been previously discussed in the academic literature are largely responsible for the increasing risk. First, the target might interpret nonnuclear attacks against its dual-use C3I assets that were motivated by conventional warfighting goals as preparations for nuclear use. It might respond to such "misinterpreted warning," to coin a term, by trying to deter the nuclear strike it believed might be coming or to mitigate its potentially calamitous consequences. Such efforts, which might include provocative nonnuclear operations to protect remaining C3I assets (such as strikes against anti-satellite weapons deep within the adversary's territory) accompanied, perhaps, by nuclear threats, could prove highly escalatory. These escalation pressures could arise even if the recipient of misinterpreted warning were not concerned about the survivability of its nuclear forces—a key distinction from crisis instability.

---

7. To the best of the author's knowledge, the first use of the term "entangled" in this general sense occurs in John D. Steinbruner, *Principles of Global Security* (Washington, D.C.: Brookings Institution Press, 2000), p. 55.

8. Avery Goldstein, "First Things First: The Pressing Danger of Crisis Instability in U.S.-China Relations," *International Security*, Vol. 37, No. 4 (Spring 2013), pp. 67–68, doi:10.1162/ISEC_a_00114; and Stephen Biddle and Ivan Oelrich, "Future Warfare in the Western Pacific: Chinese Antiaccess/Area Denial, U.S. AirSea Battle, and Command of the Commons in East Asia," *International Security*, Vol. 41, No. 1 (Summer 2016), pp. 44–45, doi:10.1162/ISEC_a_00249.

9. There are passing references to this possibility in Arbatov, Dvorkin, and Topychkanov, "Entanglement as a New Security Threat," p. 31; Zhao and Li, "The Underappreciated Risks of Entanglement," p. 51; and James N. Miller Jr. and Richard Fontaine, "A New Era in U.S.-Russian Strategic Stability: How Changing Geopolitics and Emerging Technologies Are Reshaping Pathways to Crisis and Conflict" (Cambridge, Mass. and Washington, D.C.: Belfer Center for Science and International Affairs, John F. Kennedy School of Government, Harvard University, and Center for a New American Security, September 2017), p. 19, https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-ProjectPathways-Finalb.pdf?mtime=20170918101504.

Second, a state with a damage-limitation doctrine would rely on sophisticated C3I capabilities to locate and destroy its opponent's nuclear forces and conduct missile defense operations. If these dual-use enabling capabilities were subject to attack in a conventional conflict—or even if their possessor feared they might be—the state could worry that its window of opportunity for conducting effective damage-limitation operations might have closed by the time the war turned nuclear. In this case, the state might take escalatory measures to protect its C3I system or even initiate counterforce operations preemptively. This escalation mechanism, which might be termed the "damage-limitation window," is distinct from crisis instability because it is driven by the state's desire to hold an opponent's nuclear forces at risk, not to protect its own. It is distinct from misinterpreted warning because it could operate even if the state did not believe that nuclear use by an adversary might be imminent; the state would only have to believe that such escalation was possible later on.

An additional implication of C3I entanglement is that the risks of crisis instability are more serious than portrayed in the academic literature. Scholarly warnings about crisis instability have focused on the potential for U.S. nonnuclear operations to degrade Chinese nuclear forces, but have also identified the risk of inadvertent threats to China's nuclear C3I capabilities located in the theater of operations.[10] These threats have received particular attention since the United States acknowledged, in 2013, that it seeks to defeat potential adversaries' antiaccess/area-denial capabilities by holding relevant C3I assets at risk as part of the concept formerly known as AirSea Battle (which was renamed, in 2015, as the Joint Concept for Access and Maneuver in the Global Commons and has since been further developed).[11] If overlap exists between the communication systems for China's land-based nuclear and nonnuclear missiles, as some analysts have suggested, China could mistake U.S. strikes designed to disable its nonnuclear missiles as an attack against its nuclear forces.[12]

---

10. Talmadge, "Would China Go Nuclear?" pp. 78–79, 83.
11. Air-Sea Battle Office, "Air-Sea Battle: Service Collaboration to Address Anti-Access and Area Denial Challenges" (Washington, D.C.: U.S. Department of Defense, May 2013), p. 7, http://archive.defense.gov/pubs/ASB-ConceptImplementation-Summary-May-2013.pdf.
12. Christensen, "The Meaning of the Nuclear Evolution," p. 468. For a slightly dated description of Chinese command and control that implies an overlap, see John Wilson Lewis and Xue Litai, *Imagined Enemies: China Prepares for Uncertain War* (Stanford, Calif.: Stanford University Press, 2006), pp. 197–201. For opposing views, see Cunningham and Fravel, "Assuring Assured Retaliation," pp. 42–45; and Michael Glosny, Christopher Twomey, and Ryan Jacobs, "U.S.-China Strategic Dialogue, Phase VIII Report" (Monterey, Calif.: Center on Contemporary Conflict, Naval

Entanglement, however, has created other potential triggers for crisis instability. The United States, for example, has—or could develop—incentives to launch nonnuclear kinetic attacks against existing and probable future dual-use Chinese or Russian early-warning capabilities, including over-the-horizon radars, ballistic missile early-warning radars (BMEWRs), and early-warning satellites, that are located outside the theater of operations.[13] (Kinetic weapons, which often use explosive warheads, aim to damage or destroy targets by transferring kinetic energy to them through physical contact; nonkinetic weapons include directed energy and cyber capabilities.) Moreover, Russian strikes on the United States could precipitate crisis instability if U.S. communication aircraft (currently, the United States' most survivable means to communicate with its nuclear forces) become vulnerable.

Entanglement could catalyze escalation in any major U.S.-Chinese or U.S.-Russian conventional conflict, irrespective of its origins. That said, for the sake of concreteness, the kind of U.S.-Chinese conflict that forms the backdrop to this article would most likely begin with a Chinese attempt to reunify with Taiwan by force (either unprovoked or because the government of Taiwan had declared independence), followed by U.S. intervention on behalf of Taiwan. The most probable cause of a major U.S.-Russian conflict would be the invasion and occupation of one or more of the Baltic states by Russia, followed by a U.S.-led counterattack to liberate them. In both cases, fighting could spread from the theater in which it started.

There would, of course, be important differences between the escalation dynamics in a U.S.-Chinese and U.S.-Russian conflict. Nevertheless, there would also be important similarities that help illustrate the general nature of the risks stemming from entanglement. In particular, entanglement could not only precipitate the use of nuclear weapons directly, but could also frustrate efforts to manage nonnuclear escalation, thus raising the risk of nuclear use later on. Early in a conflict, for example, to emphasize its limited war aims, the United States might refrain from conducting nonnuclear strikes beyond a certain distance into an adversary's territory. Subsequently, if the United States became worried that key C3I satellites were at risk, it might believe that it had to

Postgraduate School, November 2014), p. 10, http://calhoun.nps.edu/bitstream/handle/10945/44733/2014%20008%20-%20US-China%20Phase%20VIII%20Report.pdf.

13. There are passing references to this possibility in Arbatov, Dvorkin, and Topychkanov, "Entanglement as a New Security Threat." Over-the-horizon radars are briefly mentioned in Christopher P. Twomey, "Asia's Complex Strategic Environment: Nuclear Multipolarity and Other Dangers," *Asia Policy*, January 2011, p. 64.

attack Chinese or Russian anti-satellite (ASAT) weapons located further beyond the border.

This article begins by outlining the technological and doctrinal developments that are increasing entanglement. It then lays out three mechanisms—misinterpreted warning, the damage-limitation window, and crisis instability—by which entanglement might spark escalation and identifies the conditions under which escalation would be most likely. To provide a concrete demonstration of the severity of the escalation risks, the article then describes the likely effectiveness and effects of nonnuclear kinetic attacks against the U.S. early-warning system. It also considers the risks of cyber interference with dual-use Chinese, Russian, and U.S. early-warning assets, and in particular, the danger of the target's misinterpreting cyber espionage as an attempt to disable or destroy those assets.

With risk reduction likely to prove difficult, unilateral restraint and actions represent the most feasible policy responses for the short term. Although such steps would likely be only moderately effective in themselves, they could help pave the way for cooperative efforts in the future. Although difficult to orchestrate, cooperative risk-reduction would be desirable because, as this article emphasizes in the conclusion, the risks created by entanglement are likely to grow in the future, absent action to mitigate them.

The Technological and Doctrinal Drivers of Entanglement

Entanglement describes interactions between the nuclear and nonnuclear domains. For current purposes, its most important manifestations are the dual-use nature of many C3I assets as well as nonnuclear threats (real or perceived) to nuclear forces or their C3I infrastructure. Other manifestations, mentioned only in passing here, are dual-use delivery systems; nuclear delivery systems that are superficially similar to nonnuclear ones; and the colocation of nuclear and nonnuclear delivery systems or C3I assets. Since the end of the Cold War, entanglement has increased significantly—and, indeed, is still increasing—as the result of four trends in military technology and doctrine.

GROWING TECHNOLOGICAL THREATS

First, profound changes in weaponry have significantly magnified nonnuclear threats to states' C3I assets and, to a lesser extent, their nuclear forces. These changes include the deployment of two entirely new classes of weapons: cyberweapons (which could threaten both C3I capabilities and nuclear forces) and nonnuclear strategic ballistic missile defense systems (which could inter-

cept nuclear weapons after launch). The effectiveness of existing types of nonnuclear weapons has also improved dramatically. For example, although both the United States and the Soviet Union had some capability to target satellites without nuclear weapons by the end of the Cold War, nonnuclear ASAT weapons—both kinetic and non-kinetic—pose a much more potent threat today.[14] High-precision conventional weapons have also improved significantly, including with the introduction of satellite-guided munitions. Over the next couple of decades, further substantial improvements can be expected in all of these weapon types, and entirely new types of nonnuclear weapons, including long-range hypersonic weapons, may be deployed.[15]

growing vulnerability of c3i capabilities

Second, changes in enabling technologies have exacerbated the growing vulnerability of the C3I assets involved in nuclear operations (whether these assets are dual use or not). Digital networks have become ubiquitous, for example, creating the possibility of cyber interference. Moreover, the United States, at least in an effort to reduce costs, has pursued greater commonality in the enabling systems, such as the receivers for satellite signals, associated with different nuclear-weapon delivery systems.[16] This development, however, could magnify cyber risks. If, for example, there was a design flaw in a common receiver that left it vulnerable to being disabled by a cyberattack, then all the nuclear-weapon delivery systems that used the receiver could be simultaneously compromised.

Another cause of this growing vulnerability—at least for the U.S. nuclear C3I system—is a reduction in redundancy (there is insufficient publicly available information to assess how the redundancy of the Chinese and Russian systems has changed).[17] In the late 1980s and early 1990s, for example, two largely independent satellite-based communication systems were in use for transmitting orders for the employment of U.S. nuclear weapons.[18] The

14. Laura Grego, "A History of Anti-Satellite Programs" (Cambridge, Mass.: Union of Concerned Scientists, January 2012), http://www.ucsusa.org/sites/default/files/legacy/assets/documents/nwgs/a-history-of-ASAT-programs_lo-res.pdf.
15. Acton, "Silver Bullet?"
16. Department of the Air Force, U.S. Department of Defense, "Department of Defense Fiscal Year (FY) 2017 President's Budget Submission: Other Procurement, Air Force" (Washington, D.C.: U.S. Department of Defense, February 2016), p. 267, line item 834210, http://www.saffm.hq.af.mil/Portals/84/documents/FY17/AFD-160208-049.pdf?ver=2016-08-24-102038-590.
17. For an overview of the late 1980s system, see Peter Vincent Pry, *The Strategic Nuclear Balance*, Vol. 2: *Nuclear Wars: Exchanges and Outcomes* (New York: Crane Russak, 1990), pp. 18–22.
18. Curtis Peebles, *High Frontier: The U.S. Air Force and the Military Space Program* (Washington, D.C.: Air Force History and Museums Program, 1997), pp. 44–54, http://www.dtic.mil/dtic/tr/

Defense Satellite Communications System served intercontinental ballistic missiles (ICBMs). A separate system, the Air Force Satellite Communications System (AFSATCOM), served ICBMs, sea-launched ballistic missiles (SLBMs), and nuclear-armed aircraft, and consisted of special transponders hosted on tens of satellites mostly used for other purposes.[19] Today, the United States is in the process of deploying just four Advanced Extremely High Frequency (AEHF) satellites that will be the nation's sole space-based system for transmitting nuclear employment orders once legacy Milstar satellites have been retired. Similarly, at the end of the Cold War, the United States operated two independent networks of radio antennae to communicate with submarines.[20] One of these networks, which could provide global coverage using two extremely low-frequency antennae in the continental United States, has since been shut down.[21] Although modernization of the remaining assets would presumably enable them to function more effectively in the extraordinarily stressful conditions of a nuclear war, the overall loss of redundancy—a consequence of budgetary pressures—appears to have left the U.S. nuclear C3I system less resilient against nonnuclear attack.

growing reliance on dual-use c3i assets

Third, the U.S. nuclear C3I system has always used some dual-use assets, and is becoming increasingly reliant on them, raising the likelihood of its being attacked in a nonnuclear conflict. The United States has, for example, never fielded communication satellites that were used exclusively for nuclear operations.[22] Today, Milstar and AEHF satellites represent the United States' most secure space-based means of communicating with both nuclear and "high-priority" nonnuclear users (users tasked with particularly important or time-critical missions).[23] In fact, the vast majority of data transmitted by these

---

fulltext/u2/a442844.pdf. The systems were not entirely independent, because some transponders belonging to the Air Force Satellite Communications System were hosted by Defense Satellite Communications System satellites.

19. A 1981 estimate suggested that as many as thirty AFSATCOM transponders could be deployed by 1990. See Mark Hewish, "Satellites Show Their Warlike Face," *New Scientist*, October 1, 1981, p. 39.

20. U.S. Department of the Navy, "Submarine Communications Master Plan" (Washington, D.C.: U.S. Department of the Navy, December 1995), appendix B, http://fas.org/man/dod-101/navy/docs/scmp/part07.htm.

21. Robert Imrie, "Navy to Shut Down Sub Radio Transmitters," Associated Press, September 26, 2004, http://usatoday30.usatoday.com/tech/news/2004-09-26-sub-radio-offair_x.htm.

22. Peebles, *High Frontier*, pp. 44–52.

23. Air Force Space Command, "Advanced Extremely High Frequency System" (Washington, D.C.: U.S. Air Force, March 22, 2017), http://www.afspc.af.mil/About-Us/Fact-Sheets/Display/Article/249024/advanced-extremely-high-frequency-system/.

satellites is almost certainly associated with nonnuclear operations. Because it could be difficult for an adversary to disrupt the operation of these satellites in non-destructive ways (jamming, for example), they could become targets of direct attack in a conventional conflict.

Starting in the last decade of the Cold War, the United States has increased reliance on dual-use systems by assigning nonnuclear roles to C3I assets that used to be employed solely for nuclear operations. Until the mid-1980s, for example, U.S. early-warning satellites were used exclusively for detecting the launch of nuclear-armed missiles.[24] Today, they enable a variety of nonnuclear missions by, for example, providing cuing information for missile defenses involved in intercepting conventional ballistic missiles.[25]

In a parallel series of developments, the United States has dismantled various land-based nuclear-only communication capabilities. For example, the Emergency Rocket Communications System, which could transmit employment orders from modified ICBMs launched to overfly missile fields in the United States, was taken offline in the 1990s.[26] A decade or so later, the Survivable Low Frequency Communications System, which allowed ICBMs to receive launch orders from radio antennae, was also scrapped.[27]

The net effect of these developments is that, today, most assets in the U.S. nuclear C3I system "support both nuclear and conventional missions," according to the U.S. Government Accountability Office.[28] In fact, every C3I asset listed in the 2018 Nuclear Posture Review is known to be dual use, except for nuclear-weapon control capabilities directly associated with delivery systems (and perhaps also the United States' system for detecting nuclear

24. Their adoption for nonnuclear missions is discussed in Norman Friedman, *Seapower and Space: From the Dawn of the Missile Age to Net-Centric Warfare* (Annapolis: Naval Institute Press, 2000), pp. 242–245.
25. Committee on an Assessment of Concepts and Systems for U.S. Boost-Phase Missile Defense in Comparison to Other Alternatives and Division on Engineering and Physical Science of the National Research Council, *Making Sense of Ballistic Missile Defense: An Assessment of Concepts and Systems for U.S. Boost-Phase Missile Defense in Comparison to Other Alternatives* (Washington, D.C.: National Academies Press, 2012), p. 116, https://www.nap.edu/catalog/13189/making-sense-of-ballistic-missile-defense-an-assessment-of-concepts.
26. Federation of American Scientists, "Emergency Rocket Communications System (ERCS)" (Washington, D.C.: Federation of American Scientists, April 27, 1998), http://fas.org/nuke/guide/usa/c3i/ercs.htm.
27. Carla Williams, "Minot Completes Minuteman Emergency Communications Upgrade" (Washington, D.C.: U.S. Air Force, November 17, 2005), http://www.af.mil/News/ArticleDisplay/tabid/223/Article/132716/minot-completes-minuteman-emergency-communications-upgrade.aspx.
28. Christina Chaplain, "Nuclear Command, Control, and Communications: Update on DOD's Modernization," GAO-15-584R (Washington, D.C.: U.S. Government Accountability Office, June 15, 2015), p. 1, http://www.gao.gov/assets/680/670801.pdf.

explosions—though some of its detectors are hosted by Global Positioning System satellites).[29]

The Russian nuclear C3I system probably also includes some dual-use assets. In a 2007 edition of the journal *Military Thought*, published by the Russian ministry of defense, one retired and one serving military officer describe how satellites then under development would be used for communicating with "strategic and nonstrategic nuclear forces," as well as nonnuclear forces and even "federal and regional government agencies."[30] Their description appears to refer to communication satellites that have since been deployed as part of Russia's Unified Satellite Communication System. Separately, according to state-controlled Russian media outlets, Moscow has recently acquired a number of airborne command posts capable of communicating with both nuclear and conventional forces.[31] Moreover, as discussed below, various types of Russian radars are already dual use, and Russia's new early-warning satellites could take on nonnuclear missions in the future.

The extent of the overlap between the communication systems for China's land-based nuclear and conventional missiles has been the subject of considerable debate among analysts.[32] Beijing's recent deployment of the DF-26 ballistic missile provides some additional evidence that this overlap is significant. The warhead (or warheads) on an individual missile body can, according to an apparently authoritative Chinese source, be rapidly switched between nuclear and conventional variants.[33] This capability suggests that the physical communication infrastructure associated with these missiles can be used to transmit nuclear and nonnuclear employment orders. This evidence is not definitive, however, because it is possible that missiles are transferred between nuclear and conventional missile brigades when the warhead type is changed (though this procedure would seem to obviate the whole purpose of the "change the

29. U.S. Department of Defense, "Nuclear Posture Review," pp. 56–57.
30. V.A. Grigoryev and I.A. Khvorov, "Military Satellite Communications Systems: Current State and Development Prospects," *Military Thought*, Vol. 16, Nos. 3–4 (July 1, 2007), p. 149; see also p. 150.
31. "Russian Next-Generation 'Doomsday Plane' Finally Ready for Action," Sputnik, July 28, 2016, http://sputniknews.com/russia/20160728/1043728673/russia-doomsday-plane-ready.html.
32. Christensen, "The Meaning of the Nuclear Evolution," p. 468; Lewis and Xue, *Imagined Enemies*, pp. 197–201; Cunningham and Fravel, "Assuring Assured Retaliation," pp. 42–45; and Glosny, Twomey, and Jacobs, "U.S.-China Strategic Dialogue, Phase VIII Report," p. 10.
33. Andrew S. Erickson, "Academy of Military Science Researchers: 'Why We Had to Develop the Dongfeng-26 Ballistic Missile'—Bilingual Text, Analysis, and Related Links," andrewerickson.com, December 5, 2015, http://www.andrewerickson.com/2015/12/academy-of-military-science-researchers-why-we-had-to-develop-the-dongfeng-26-ballistic-missile-bilingual-text-analysis-links/.

warhead, not the missile" capability).[34] Additionally, as discussed below, various Chinese early-warning capabilities are already, or may become, dual use.

growing doctrinal threats

Fourth, the military doctrines of China, Russia, and the United States appear to envision attacks on space- and land-based C3I assets, including dual-use ones, to further conventional warfighting goals. In the case of the United States, this tactic was explicitly articulated in the AirSea Battle concept. Meanwhile, Washington has openly expressed concern that both China and Russia seek to hold U.S. C3I satellites at risk to support potential efforts to undermine U.S. conventional operations.[35] The U.S. intelligence community has highlighted the threat from both states to U.S. early-warning satellites, in particular.[36] A consistent picture is painted by Chinese and Russian sources. For example, the *Science of Second Artillery Campaigns*, a classified but leaked textbook from 2004 believed to contain an authoritative description of China's strategic doctrine, appears to endorse attacks against U.S. early-warning radars as a way of suppressing missile defenses in a conventional conflict.[37] Moreover, Chinese experts have openly advocated for the ability to attack U.S. early-warning satellites.[38] In a similar vein, Russian experts have stated that, in a conventional conflict, Moscow would consider attacking U.S. C3I assets, including ground-based early-warning radars.[39]

---

34. Jordan Wilson, "China's Expanding Ability to Conduct Conventional Missile Strikes on Guam" (Washington, D.C.: U.S.-China Economic and Security Review Commission, May 10, 2016), p. 8, https://www.uscc.gov/sites/default/files/Research/Staff%20Report_China%27s%20Expanding%20Ability%20to%20Conduct%20Conventional%20Missile%20Strikes%20on%20Guam.pdf.

35. Defense Intelligence Agency, "Russia Military Power: Building a Military to Support Great Power Aspirations," DIA-11-1207-161 (Washington, D.C.: Defense Intelligence Agency, 2017), p. 36, http://www.dia.mil/Portals/27/Documents/News/Military%20Power%20Publications/Russia%20Military%20Power%20Report%202017.pdf; and Office of the Secretary of Defense, "Military and Security Developments Involving the People's Republic of China 2017," annual report to Congress (Washington, D.C.: U.S. Department of Defense, 2017), p. 35, https://www.defense.gov/Portals/1/Documents/pubs/2017_China_Military_Power_Report.PDF?ver=2017-06-06-141328-770.

36. Daniel R. Coates, "Worldwide Threat Assessment of the U.S. Intelligence Community," statement for the record (Washington, D.C.: Office of the Director of National Intelligence, March 6, 2018), p. 13, https://www.dni.gov/files/documents/Newsroom/Testimonies/Final-2018-ATA---Unclassified---SASC.pdf.

37. Second Artillery Corps, People's Liberation Army, *The Science of Second Artillery Campaigns*, unclassified U.S. government translation (Beijing: PLA Press, 2004), pp. 397–398. Given that this section discusses suppressing both missile and air defenses, this reference is probably to both missile and aircraft early-warning radars.

38. Zhao and Li, "The Underappreciated Risks of Entanglement," p. 51; and Chase, Erickson, and Yeaw, "Chinese Theater and Strategic Missile Force Modernization and Its Implications for the United States," p. 83.

39. Arbatov, Dvorkin, and Topychkanov, "Entanglement as a New Security Threat," p. 31.

Escalation Pathways: Effects of Entanglement on Conflict Dynamics

One consequence of growing entanglement is the possibility of "incidental attacks" on an opponent's nuclear forces or their enabling capabilities. In such an attack, one state strikes an adversary's dual-use assets to influence the outcome of a conventional conflict but, in the process, inadvertently degrades its nuclear capabilities.[40] Strikes against dual-use C3I capabilities—communication and early-warning assets, in particular—would probably represent the most consequential type of incidental attack. Incidental attacks could also result, however, from strikes against dual-use weapon delivery platforms, such as aircraft and missiles.

Incidental attacks have the potential to be escalatory, in no small part because it could be effectively impossible for the target to distinguish them from deliberate attacks intended to undermine its ability to conduct nuclear operations (including obtaining warning of an incoming nuclear strike). The general difficulty of assessing intent would likely be compounded by the fog of war, which would probably be thick in any major conventional conflict and further exacerbated by likely attacks against ISR capabilities. Moreover, as Barry Posen argued, a variant of the security dilemma might arise: prudence could require a state to treat attacks on its nuclear forces or their enabling capabilities as deliberate and take actions to protect them; to assume that surviving assets were not threatened would carry the risk that they might be destroyed if the enemy's intent had been misjudged.[41]

There are three distinct pathways—misinterpreted warning, the damage-limitation window, and crisis instability—through which actual or threatened incidental attacks could spark inadvertent escalation.

MISINTERPRETED WARNING

In a conventional war between two nuclear-armed states, nonnuclear attacks against an opponent's dual-use enabling capabilities motivated by conventional warfighting goals could be indistinguishable from operations intended to prepare the battlespace for nuclear use. Such attacks, therefore, could create misinterpreted warning—especially if the state launching them was in danger of losing the war.

Although a state concerned about becoming the target of a nuclear attack might not use nuclear weapons immediately, its concern might lead it to act in ways that could catalyze further escalation, raising the likelihood of nuclear

40. This definition is somewhat different from the one in Posen, *Inadvertent Escalation*, p. 2.
41. Ibid., pp. 12–16.

use later on. The state would be motivated by a desire to avoid or mitigate the potentially catastrophic costs of becoming the target of even a limited nuclear strike; in contrast to crisis instability, these escalation pressures could be felt even if the state was not concerned about its nuclear forces being vulnerable or its ability to transmit employment orders to them.

Two questions arise when assessing the escalation risks of misinterpreted warning. First, how likely is it that the target would interpret nonnuclear strikes against its dual-use C3I assets as possible preparations for nuclear use? Second, if the target did become concerned that it might shortly be on the receiving end a nuclear strike, would it be likely to react in ways that tended to catalyze further escalation?

Because Moscow and Beijing have different nuclear postures and doctrines, there are somewhat different reasons why their striking dual-use U.S. enabling assets might generate misinterpreted warning. U.S. incidental strikes on dual-use Chinese or Russian C3I assets could also lead to misinterpreted warning, though this possibility is not discussed further here.

HOW MISINTERPRETED WARNING COULD OCCUR. The United States government has indicated its belief that, in a conventional conflict, Russia might opt for limited nuclear use in an attempt to compel the United States into backing down—a strategy sometimes termed "escalate to de-escalate" in the Western discourse.[42] It also appears to worry that, if a limited nuclear war escalated, Russia might launch large-scale damage-limitation strikes against U.S. nuclear forces (even though such strikes could not deprive the United States of a second-strike capability today).[43] Whether these beliefs accurately reflect Russian strategy is essentially immaterial for current purposes; rather, they are important because, right or wrong, they would likely inform the United States' assessment of Russia's intentions in a conflict. In this way, for at least three reasons, these beliefs create the potential for Washington to misinterpret Russian incidental strikes against dual-use U.S. C3I assets as preparations for nuclear use.

First, Russia might attack ground-based or space-based U.S. early-warning assets to defeat European missile defenses that were proving effective in inter-

42. U.S. Department of Defense, "Nuclear Posture Review," p. 30; and Robert Work and James Winnefeld, prepared statement, *Nuclear Deterrence in the 21st Century*, hearing before the Committee on Armed Services, U.S. House of Representatives, 114th Cong., 1st sess., June 25, 2015, p. 4, http://docs.house.gov/meetings/AS/AS00/20150625/103669/HHRG-114-AS00-Wstate-WinnefeldJrUSNJ-20150625.pdf.

43. The emphasis that the United States places on force survivability in official policy can only be explained by concerns about damage-limiting Russian strikes.

cepting its nonnuclear missiles. Washington might see such attacks, however, as preparations to ensure that limited nuclear strikes by Russia could penetrate the United States' homeland missile defenses. Government-affiliated Russian experts have publicly advocated "limited strategic strikes" against the U.S. homeland under a variety of circumstances (including if Russia became concerned that the United States was about to embark on a conventional counterforce campaign against its nuclear forces).[44] Such experts have also expressed concern that U.S. missile defenses might be capable of defeating such strikes. Indeed, the United States has declared that homeland defenses "would be employed to defend the United States against limited missile launches from any source" (even if such defenses cannot cope with large-scale attacks).[45] In response, Russian strategists have suggested that, prior to launching limited strategic strikes, Moscow should try to neutralize those defenses by attacking the U.S. early-warning system.[46] If Washington interpreted strikes against its early-warning capabilities in this light, misinterpreted warning could arise.

Second, Russia could attack dual-use U.S. communication assets to undermine a variety of American nonnuclear operations. Washington could interpret such attacks, however, as an attempt to forestall a proportionate U.S. response to the limited use of low-yield nuclear weapons. Nuclear-armed aircraft might well be the United States' preferred means of responding to a limited nuclear strike, because the B-61 gravity bomb has the lowest-yield nuclear option in the U.S. arsenal.[47] The communication links for deployed aircraft, however, are particularly vulnerable to being severed.[48] Russian incidental strikes might destroy the satellites and ground-based transmitters that could enable communications with aircraft operating over or around Russia. Meanwhile, communication aircraft operating over the United States would probably be too distant to direct operations in that region. Washington, therefore, could interpret Russian attacks against U.S. communication links as an attempt to deny the United States the ability to respond in kind to a low-yield

44. Arbatov, Dvorkin, and Topychkanov, "Entanglement as a New Security Threat," pp. 20–21.
45. U.S. Department of Defense, "Ballistic Missile Defense Review Report" (Washington, D.C.: U.S. Department of Defense, February 2010), p. 13, http://archive.defense.gov/bmdr/docs/BMDR%20as%20of%2026JAN10%200630_for%20web.pdf.
46. Arbatov, Dvorkin, and Topychkanov, "Entanglement as a New Security Threat," p. 31.
47. The 2018 Nuclear Posture Review calls for the acquisition of additional submarine-based low-yield nuclear capabilities. Whether and when these weapons will be deployed remains to be seen. See U.S. Department of Defense, "Nuclear Posture Review," pp. 54–55.
48. In theory, the United States could still use aircraft for nuclear operations by "pre-programming" targets at take-off or shortly afterward. This approach, however, would undermine the maintenance of positive control throughout a flight, which is a key rationale for maintaining nuclear-armed aircraft.

nuclear strike in the hope that it would be deterred from a more forceful response by the fear of further escalation.

Third, Russian attacks against dual-use U.S. early-warning or communication assets would risk being seen as a signal of Russia's resolve to use nuclear weapons unless the United States conceded to its demands. In an effort to deter limited nuclear use by Russia, senior U.S. officials have publicly stressed the risk of escalation to a strategic nuclear war, stating, for example, that "anyone who thinks they can control escalation through the use of nuclear weapons is literally playing with fire."[49] Because Russian incidental strikes against dual-use U.S. C3I assets could help Russia fight a strategic nuclear war, they could be interpreted by Washington as an effort to enhance the credibility of limited nuclear use. For example, degrading the U.S. early-warning system might prevent the United States from launching ICBMs, dispersing bombers, or sheltering national leaders before they were eliminated in a nuclear attack. Similarly, disabling communication systems might slow a U.S. nuclear response to a Russian counterforce strike, giving Russia time for follow-up damage-limitation strikes.

To be sure, the United States' interpretation of Russian strikes against dual-use U.S. enabling assets would likely depend on the context. Had Russia raised the alert level of its nuclear forces, dispersed them, or even issued orders to prepare them for nuclear employment? Had Moscow put into action plans to try to ensure the continuity of government in the event of a nuclear war? What messages was the government sending to its own population? Would it be threatened from within if it lost the war? In practice, such questions could be extremely difficult to answer because, by the time that Russia had attacked dual-use U.S. early-warning and communication assets, it would probably have launched extensive attacks against U.S. ISR capabilities, potentially denying much needed contextual information to the United States.[50] In the absence of this information, Washington might feel its most prudent course of action was to assume the worst about Moscow's intentions.

The risk of the United States' misinterpreting Chinese nonnuclear strikes against dual-use U.S. C3I assets as preparations for nuclear use would probably be lower than in the case of Russia for two reasons. First, in contrast to Moscow, Beijing has adopted a no-first-use pledge. Second, unlike their

49. Work and Winnefeld, prepared statement, p. 4. See also U.S. Department of Defense, "Nuclear Posture Review," p. 30.
50. Forrest E. Morgan, "Deterrence and First-Strike Stability in Space: A Preliminary Assessment" (Santa Monica, Calif.: RAND Corporation, 2010), p. 19, http://www.rand.org/content/dam/rand/pubs/monographs/2010/RAND_MG916.pdf.

Russian counterparts, Chinese leaders can have absolutely no doubt that nuclear first use would do nothing to meaningfully limit the damage their country would suffer in a nuclear war with the United States. As a result, Washington would be unlikely to interpret Chinese nonnuclear strikes as preparations to fight and win a strategic nuclear war.

That said, the United States could still interpret Chinese attacks against its early-warning system as preparations for limited nuclear strikes intended to terrify the United States into terminating a conflict on terms not too unfavorable to Beijing. Fairly or not, Washington does not have complete confidence in the reliability of China's no-first-use pledge.[51] In particular, skeptics typically argue that Beijing would be most likely to abandon this pledge if China were in danger of losing a war over Taiwan—an outcome that could jeopardize the continued rule of the Chinese Communist Party.[52] If, in this circumstance, China attacked critical U.S. early-warning assets—satellites, in particular—in an effort to help its conventional ballistic missiles penetrate U.S. defenses, Washington might conclude that desperate Chinese leaders were preparing limited nuclear strikes, against either the United States or regional targets.[53]

Again, much would depend on context. The likelihood of misinterpreted warning would probably increase if, in addition to attacking dual-use U.S. enabling capabilities, China had dispersed or alerted nuclear-armed missiles. Although this step could be a standard defensive precaution to protect the missiles' survivability in a major conflict, it might also exacerbate concerns in Washington about the possibility of Chinese first use. Some nuclear-armed medium-range DF-21A ballistic missiles appear to be targeting U.S. assets in the West Pacific.[54] The alerting of these missiles could be seen by the United States, therefore, as preparations for regional nuclear strikes. The alerting of

51. These concerns are strongly suggested, although not stated explicitly, in U.S. Department of Defense, "Nuclear Posture Review," p. 32.

52. Mark Schneider, "The Nuclear Doctrine and Forces of the People's Republic of China" (Fairfax, Va.: National Institute Press, November 2007), pp. 7–8, http://www.nipp.org/wp-content/uploads/2014/12/China-nuclear-final-pub.pdf.

53. Early in a conflict, probable Chinese attacks on regional missile-defense radars would likely be less escalatory, because China would probably not be facing defeat then and because such radars are not critical to nuclear operations. The PAVE PAWS radar in Taiwan is a special case and discussed below.

54. Because certain Chinese missile brigades (notably, the 807 brigade in Anhui, but also perhaps the 810 brigade in Liaoning and the 816 brigade in Jilin) are not within range of important Russian or Indian targets, it is difficult to see what other role they could serve. See Jeffrey Lewis, *Paper Tigers: China's Nuclear Posture*, Adelphi 446 (Abingdon, U.K.: Routledge for the International Institute for Strategic Studies, 2014), p. 116. See also U.S. Department of Defense, "Nuclear Posture Review," p. 31.

China's ICBM force, meanwhile, could be interpreted as an attempt to threaten the U.S. homeland and so deter nuclear retaliation to Chinese first use against regional targets. The escalation pressures might be more serious still if China had conducted extensive attacks against U.S. ISR assets, denying the United States contextual information that might be helpful in interpreting Chinese intentions correctly.

HOW THE UNITED STATES MIGHT RESPOND TO MISINTERPRETED WARNING. The United States' response to misinterpreted warning would probably depend on a range of factors, including its assessment of the likelihood of nuclear use by the adversary. Nonetheless, an overriding consideration would probably be to deter such use or, if deterrence failed, to limit the damage that the United States would suffer in a nuclear war—a goal explicitly articulated in the 2018 Nuclear Posture Review.[55] As such, misinterpreted warning could lead to at least three general types of U.S. response; none of which is mutually exclusive and all of which could spark further escalation.

First and most immediate, the United States would probably seek to protect surviving elements of its nuclear C3I system because of their importance to damage-limitation efforts, including counterforce attacks and missile defense operations. As described below, for these efforts to have any hope of success, the United States would have to preserve much more than just the relatively basic capability needed to transmit employment orders to survivable nuclear forces. Steps to preserve surviving C3I capabilities could prove escalatory. For example, the United States might attack ASAT weapons that it believed could threaten important U.S. satellites. If these weapons were located deep inside China or Russia, then such attacks could spark escalation, especially if the United States had previously avoided striking far inside its adversary's borders in an effort to keep the war limited. Alternatively, or additionally, the United States could launch tit-for-tat strikes against equivalent Chinese or Russian enabling assets in an attempt to coerce Beijing or Moscow into ceasing attacks on U.S. C3I assets—potentially leading the adversary to fear for the survivability of its nuclear forces and generating crisis instability.

Second, misinterpreted warning might prompt the United States to alert bombers and send additional ballistic missile submarines (SSBNs) to sea. Although neither China nor Russia could hope to disarm the United States, both could plausibly threaten U.S. submarines in port and bombers at their bases. In consequence, enhancing the survivability of these platforms might seem

55. U.S. Department of Defense, "Nuclear Posture Review," p. 23.

to Washington like a sensible precaution. If the adversary were not planning to use nuclear weapons, however, this precaution could appear to be threatening. In particular, Beijing or Moscow might worry about the possibility of attacks with very short warning times launched from forward-deployed stealthy bombers or from SSBNs firing SLBMs on depressed trajectories from near its coasts. In turn, China or Russia might respond by taking steps to enhance the survivability of its nuclear forces, such as dispersing mobile missiles, which could appear to confirm Washington's fears. In this way, misinterpreted warning and crisis instability could exacerbate each other.

Third, the United States could threaten to use—or even use—nuclear weapons in response to misinterpreted warning. Following attacks on dual-use U.S. C3I assets, Washington might threaten to use nuclear weapons if the attacks continued or if the adversary employed nuclear weapons. Such a threat, however, could trigger an escalation cycle similar to the one that might be sparked by the dispersal of U.S. SSBNs and bombers. Alternatively, if the adversary did not judge the threat to be credible and continued to attack U.S. C3I assets, the United States might feel compelled to follow through on its threat and resort to nuclear use. Although it could attempt a disarming first strike, the limited use of nuclear weapons would probably be more likely. U.S. leaders—mirroring the precise logic that they were ascribing to their Chinese or Russian counterparts—might hope that such strikes would terrify the adversary into complying with U.S. demands.

It is even possible that the United States would respond directly to attacks on dual-use C3I assets with the use of nuclear weapons, without first issuing a nuclear threat. Although such a response would be disproportionate and thus unlikely, Washington might feel that having threatened, in the 2018 Nuclear Posture Review, to use nuclear weapons in this eventuality, it had to follow through or else risk damaging its credibility and very undermining other elements of U.S. declaratory policy.[56]

THE DAMAGE-LIMITATION WINDOW

Although all nuclear operations require C3I capabilities, the enabling requirements for damage-limitation operations would be particularly demanding. Significant damage to a state's nuclear C3I system would preclude any possi-

56. For an analysis of why immediate nuclear use would be disproportionate, see James Acton, "Command and Control in the Nuclear Posture Review: Right Problem, Wrong Solution," *War on the Rocks*, February 5, 2018, https://warontherocks.com/2018/02/command-and-control-in-the-nuclear-posture-review-right-problem-wrong-solution/.

bility of such operations being effective. Because many enabling capabilities are dual use and could be attacked or threatened in a conventional war, a state with a damage-limitation doctrine might conclude that it had only a narrow window of opportunity near the start of a conflict in which it could realistically try to attack its opponent's nuclear forces and to defend against whatever it failed to destroy. Fear that this damage-limitation window might close could create pressures for the state to conduct counterforce strikes preemptively or, more likely, initiate aggressive military operations to try to preserve the option of conducting damage-limitation operations later on.[57]

These escalation pressures differ from those created by crisis instability in that escalation would be motivated by the goal of holding an adversary's nuclear forces at risk, not of ensuring the survivability of the state's own forces. Although there are some similarities between the damage-limitation window and misinterpreted warning—particularly in that they might spark aggressive efforts to protect surviving C3I capabilities—there is one critical difference. Escalation driven by fear of the damage-limitation window's closing stems from the unavoidable possibility that a war between two nuclear-armed states might ultimately turn nuclear and could be felt even if neither state believed its adversary was currently preparing for nuclear use.

Damage-limitation operations would consist of counterforce attacks, backed up by missile defenses. The United States openly acknowledges that it plans for counterforce attacks. Specifically, according to the 2013 "Report on Nuclear Employment Strategy of the United States," the most recent authoritative public statement on U.S. targeting policy, "guidance requires the United States to maintain significant counterforce capabilities against potential adversaries."[58] Meanwhile, Washington appears to assume that Moscow might also launch counterforce attacks. By contrast, there is no evidence that Beijing contemplates such attacks, not least because it lacks the capability to conduct them on a meaningful scale. Fear of the damage-limitation window's closing, therefore, could generate escalation pressures on Russia but not China—though this section again focuses on the United States.

57. Charles L. Glaser and Steve Fetter observe, in a parallel line of reasoning, that the possibility of an adversary alerting its nuclear forces could create escalation pressures by complicating damage-limitation operations. See Glaser and Fetter, "Should the United States Reject MAD? Damage Limitation and U.S. Nuclear Strategy toward China," *International Security*, Vol. 41, No. 1 (Summer 2016), pp. 61–62, doi:10.1162/ISEC_a_00248.

58. U.S. Department of Defense, "Report on Nuclear Employment Strategy of the United States Specified in Section 491 of 10 U.S.C." (Washington, D.C.: U.S. Department of Defense, June 2013), p. 4, http://www.defense.gov/Portals/1/Documents/pubs/ReporttoCongressonUSNuclearEmploymentStrategy_Section491.pdf.

The importance of C3I capabilities to damage-limitation operations is difficult to overstate. Attacking dispersed mobile missiles would be particularly challenging. Despite much debate among U.S. strategists about how effective such efforts might prove, there is no disagreement that without high-quality ISR to detect and track missiles, along with fast and reliable communications to relay targeting data, they would be certain to fail.[59] Anti-submarine warfare operations by the United States against an enemy's SSBNs would also benefit from sophisticated enabling capabilities. Such efforts would be more likely to succeed if the operations of U.S. aircraft, surface ships, and attack submarines were coordinated, and if these platforms could share information, placing a premium on high-bandwidth communications. Meanwhile, early-warning capabilities would enable U.S. ICBMs to be launched before they were destroyed by a large-scale Russian nuclear strike (potentially enabling the United States to target any nuclear forces held in reserve by Russia). Early-warning capabilities would also be important because of their role in both regional and homeland missile defense operations. Interestingly, in this way, the existence of missile defenses is not guaranteed to reduce time pressure on the United States to act, but can, in some circumstances, actually increase it.

Even during the Cold War, when many enabling capabilities were reserved exclusively for nuclear operations and were largely invulnerable to an adversary's nonnuclear weapons, there was concern that nuclear threats to C3I assets could create escalation pressures by threatening to preclude damage limitation.[60] Today, this escalation risk is magnified by the possibility that such assets could be degraded, through incidental attacks, over the course of a conventional war.

The possibility of U.S. damage-limitation operations becoming infeasible could spark serious concern in Washington. In extremis, the United States might respond by launching counterforce attacks preemptively, while its C3I capabilities were still intact. In less extreme circumstances, it might initiate

59. For recent contributions to this debate see, for example, Glaser and Fetter, "Should the United States Reject MAD?" pp. 63–70; Austin Long and Brendan Rittenhouse Green, "Stalking the Secure Second Strike: Intelligence, Counterforce, and Nuclear Strategy," *Journal of Strategic Studies*, Vol. 38, Nos. 1–2 (2015), pp. 38–73, doi:10.1080/01402390.2014.958150; and Keir A. Lieber and Daryl G. Press, "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence," *International Security*, Vol. 41, No. 4 (Spring 2017), pp. 9–49, https://doi.org/10.1162/ISEC_a_00273.

60. Robert Jervis, *The Meaning of the Nuclear Revolution: Statecraft and the Prospect of Armageddon* (Ithaca, N.Y.: Cornell University Press, 1989), p. 165. For a Cold War analysis of how attacks on C3I capabilities could blunt the effectiveness of nuclear operations, see Ashton B. Carter, "Assessing Command System Vulnerability," in Carter, John D. Steinbruner, and Charles A. Zraket, eds., *Managing Nuclear Operations* (Washington, D.C.: Brookings Institution Press, 1987), pp. 555–610.

escalatory military operations, such as those described above, to protect these capabilities and hence preserve the option of conducting counterforce operations at a later time. As with misinterpreted warning, the United States could also threaten that further attacks against key U.S. C3I capabilities would precipitate a nuclear response. If attacks continued, it might follow through on this threat.

CRISIS INSTABILITY

Crisis instability could be induced by threats to the survival of a state's nuclear forces or their enabling capabilities.[61] In assessing the significance of such threats, the "key question," argues Caitlyn Talmadge, "would not be whether the target state expected to suffer complete nuclear disarmament . . . [but whether it] feared the erosion of its nuclear capabilities past some threshold considered vital to its security."[62] (In the general political science literature, the term "crisis instability" is often used in a somewhat different sense to describe the tendency to resort to the use of force in a crisis.)

In the Cold War, analysts generally assumed that if crisis instability led to nuclear first use, such use would be in the form of a large-scale preemptive first strike. Today, if the United States or, much more likely, Russia felt its nuclear forces or associated C3I capabilities to be in severe danger (whether from nuclear or nonnuclear threats), it might conceivably launch such a strike. Other responses, however, would probably be much more likely—including by China, which lacks the capability for effective large-scale preemption.[63] For example, a state might enhance the survivability of its nuclear forces by dispersing mobile weapons. The national leadership might pre-delegate nuclear launch authority to field commanders. To try to scare its opponent into backing down from threatening its nuclear forces, a state might threaten to use nuclear weapons or even use them in a limited way.[64] All of these steps could spark further escalation, albeit with varying likelihoods.

The possibility that nonnuclear operations might induce crisis instability was first discussed by scholars toward the end of the Cold War, in part because

61. The literature on crisis stability is vast, but the seminal discussion is Thomas C. Schelling, *The Strategy of Conflict* (Cambridge, Mass.: Harvard University Press, 1960), chap. 9. For the concept's historical origins, see Michael S. Gerson, "The Origins of Strategic Stability: The United States and the Threat of Surprise Attack," in Elbridge A. Colby and Gerson, eds., *Strategic Stability: Contending Interpretations* (Carlisle, Pa.: U.S. Army War College Press, 2013), chap. 1, http://www.strategicstudiesinstitute.army.mil/pubs/download.cfm?q=1144.
62. Talmadge, "Would China Go Nuclear?" p. 63; see pp. 57–64 more generally.
63. Michael S. Gerson, "No First Use: The Next Step for U.S. Nuclear Policy," *International Security*, Vol. 35, No. 2 (Fall 2010), pp. 35–39, doi:10.1162/ISEC_a_00018.
64. Talmadge, "Would China Go Nuclear?" pp. 58–59.

of the potential for such operations to degrade C3I capabilities. Most significantly, in his 1991 study, *Inadvertent Escalation*, Posen argued that, as the Soviet early-warning network was degraded over the course of a conventional war in Europe, Moscow might come to believe that the United States was about to decapitate the Soviet nuclear C3I system and launch a preemptive first strike.[65] At about the same time, Bruce Blair identified the vulnerability of the U.S. nuclear C3I system to Soviet nonnuclear weapons as another potential trigger of crisis instability.[66]

More recently, scholarly discussions of the implications of C3I vulnerability for crisis instability have focused on the possibility of U.S. nonnuclear attacks on Chinese C3I capabilities located in the theater of operations—in particular, the communication system for China's land-based mobile missiles, but also its air-defense radars.[67] Other C3I assets, including Chinese and Russian early-warning capabilities and U.S. communication capabilities, are also entangled, creating escalation risks that have not been identified before in the academic literature.

Russia has developed various capabilities to provide early warning of an incoming attack with nuclear-armed ballistic missiles. China, meanwhile, appears to be in the process of doing so. One potential purpose of such capabilities is to enable a state to launch nuclear weapons before they are destroyed. Russian nuclear doctrine is generally believed to include this option, known as "launch under attack" or "launch on warning" (neither term has a universally accepted definition, although the United States adopts the former in describing its own policy). There is evidence, including in the 2013 edition of *Science of Military Strategy*, a textbook published by the People's Liberation Army Academy of Military Sciences, that China may be moving in the same direction (though if so, it may be planning to alert its forces in a crisis rather than keep them on day-to-day alert).[68] Separately, both states have extensive air-

65. Posen, *Inadvertent Escalation*, chaps. 2–3. See also Bruce G. Blair, *The Logic of Accidental Nuclear War* (Washington, D.C.: Brookings Institution Press, 1993), pp. 270–271.
66. Bruce G. Blair, *Strategic Command and Control: Redefining the Nuclear Threat* (Washington, D.C.: Brookings Institution, 1985), pp. 207, 296–297; and Bruce G. Blair, "Alerting in Crisis and Conventional War," in Carter, Steinbruner, and Zraket, *Managing Nuclear Operations*, pp. 107–108.
67. On attacks against communication assets, see Chase, Erickson, and Yeaw, "Chinese Theater and Strategic Missile Force Modernization and Its Implications for the United States," pp. 105–106; Pollack, "Emerging Strategic Dilemmas in U.S.-Chinese Relations," pp. 57–58; Christensen, "The Meaning of the Nuclear Evolution," p. 468; Cunningham and Fravel, "Assuring Assured Retaliation," pp. 42, 44; and Talmadge, "Would China Go Nuclear?" pp. 78–79. On attacks against air defenses, see Talmadge, "Would China Go Nuclear?" pp. 78–79." On attacks against air defenses, see Talmadge, "Would China Go Nuclear?" pp. 77–78.
68. Gregory Kulacki, "The Chinese Military Updates China's Nuclear Strategy" (Cambridge, Mass.: Union of Concerned Scientists, March 2015), p. 4, http://www.ucsusa.org/sites/default/files/attach/2015/03/chinese-nuclear-strategy-full-report.pdf.

defense systems, which probably play an important role in protecting their nuclear forces and associated C3I capabilities from the perceived threat of nuclear or nonnuclear attack by U.S. aircraft and cruise missiles.

At least three types of Chinese or Russian early-warning assets are already entangled—or could become entangled—and, therefore, might be subject to incidental attacks by the United States, with the consequent risk of crisis instability. First, the United States might target China's or Russia's small collection of over-the-horizon radars, which can detect some threats at much greater distances than conventional line-of-sight radars.[69] As other analysts have noted, the United States might have a variety of incentives, in a conventional conflict, to strike these radars—especially, perhaps, Chinese ones with a role in locating U.S. aircraft carriers.[70] What has not been noted before (at least in the context of a discussion of escalation risks) is that China and Russia appear to regard their over-the-horizon radars as perhaps their best means of gaining at least some warning of a U.S. attack with stealthy aircraft or cruise missiles, which they worry pose a serious threat to the survivability of their nuclear forces.[71] The loss of these radars, therefore, could be particularly disquieting to Beijing or Moscow.

Second, an even more serious escalation risk that appears to have gone entirely unnoticed by analysts is incidental attacks on BMEWRs—particularly the network of these radars that rings Russia. These dual-use radars are probably Russia's most important assets for space situational awareness up to a few thousand kilometers in altitude and so enable Russia to hold numerous U.S. satellites at risk.[72] In consequence, the United States could strike this network

69. Pavel Podvig, "Russia Begins Deployment of Over-the-Horizon Radars," *Russian Strategic Nuclear Forces* blog, December 3, 2013, http://russianforces.org/blog/2013/12/russia_begins_deployment_of_ov.shtml; and Office of the Secretary of Defense, "Military and Security Developments Involving the People's Republic of China 2014," annual report to Congress (Washington, D.C.: U.S. Department of Defense, 2014) pp. 40, 69, http://www.defense.gov/Portals/1/Documents/pubs/2014_DoD_China_Report.pdf.

70. Twomey, "Asia's Complex Strategic Environment," p. 64. Of the two types of over-the-horizon radars, skywave and groundwave, the former could have a role in detecting both ships and air-breathing threats.

71. Zhou Wanxing, "Tianbo Chaoshiju Leida Fazhan Zongshu" [Summary of the development of Skywave over-the-horizon radar], *Journal of Electronics*, Vol. 39, No. 6 (2011), pp. 1375–1376 (in Chinese; the author thanks Tong Zhao for translating the relevant section of this article); Podvig, "Russia Begins Deployment of Over-the-Horizon Radars"; and "I See You: Russian-Made Sunflower Radar Is Capable of Detecting F-35 Jets," Sputnik, July 2, 2016, http://sputniknews.com/science/20160702/1042341025/russia-podsolnukh-radar-f35.html.

72. Pavel Podvig, "Status of the Russian Early-Warning Radar Network," *Russian Strategic Nuclear Forces* blog, January 13, 2013, http://russianforces.org/blog/2013/01/status_of_the_russian_early-warning.shtml.

in an effort to protect its satellites. Such attacks could generate severe crisis instability, given Russia's reliance on launch under attack.

At least two Chinese BMEWRs can be identified from publicly available satellite imagery—although it is unclear how many such radars China possesses or how many it ultimately intends to construct.[73] Chinese BMEWRs have an inherent capability to contribute to space situational awareness and hence enable ASAT operations, making them potential U.S. targets. Moreover, China may be building BMEWRs to enable the switch to a launch-under-attack posture. If it does so, China could view U.S. strikes against those radars as an attempt to undermine the survivability of its nuclear forces.

Other technological developments could exacerbate the escalation risks associated with attacks on BMEWRs yet further. Today, Chinese and Russian BMEWRs would be generally incapable of tracking most U.S. nonnuclear weapons, such as aircraft and cruise missiles (not least because of the relatively low altitude at which such weapons fly). The United States, however, is considering acquiring long-range nonnuclear ballistic missiles, which could be tracked by BMEWRs.[74] If the United States decides to deploy nonnuclear ballistic missiles, it might attack such radars, in a conflict, to suppress Chinese or Russian defenses.

Third, for a similar reason, U.S. strikes against Russian or possible Chinese early-warning satellites, which seem unlikely today, could become more plausible in the future. Since November 2015, Russia has deployed two satellites as part of a new space-based early-warning system, and it has ambitious plans to deploy "about ten" by 2020.[75] Even if such plans are only partially realized, Russia may significantly increase its reliance on space-based early warning. Meanwhile, the U.S. Department of Defense assesses that China also has an interest in acquiring early-warning satellites.[76] In fact, according to me-

73. These were identified by Catherine Dill and are located at 46.528085°N, 130.755181°E (in Heilongjiang) and 30.286637°N, 119.128591°E (in Zhejiang). Given its location, a similar radar at 41.641422°N, 86.237161°E (in Xinjiang) is probably used for monitoring China's own testing activities. Author's personal communications with Catherine Dill and Jeffrey Lewis, 2016–2018.

74. A requirement in the Fiscal Year 2018 U.S. National Defense Authorization Act is likely to involve the Department of Defense studying the feasibility of converting missile-defense interceptors into land-attack ballistic missiles. See *National Defense Authorization Act for Fiscal Year 2018*, Public Law 115-91, 115th Cong., 1st sess. (December 12, 2017), sec. 1243.(c).(2).

75. William Graham, "Soyuz 2-1B Launches Tundra Missile Detection Spacecraft," nasaspaceflight.com, May 25, 2017, https://www.nasaspaceflight.com/2017/05/soyuz-2-1b-launches-tundra-missile-detection-spacecraft/; and "Russia to Launch Ten Missile Attack Warning Satellites by 2020," TASS, December 20, 2016, http://tass.com/defense/920880.

76. Office of the Secretary of Defense, "Military and Security Developments Involving the People's Republic of China 2017," p. 61.

dia reports China had developed plans, by as early as 2014, to deploy its first such satellite.[77] For the time being, Russian and possible Chinese satellites may not contribute enough to nonnuclear military operations for them to become plausible targets of incidental U.S. strikes. If, however, the United States deploys nonnuclear ballistic missiles or hypersonic boost-glide weapons, which such satellites could track, that calculus could change, creating additional potential triggers of crisis instability.[78]

Even more dramatically, over the next decade or two, actual or threatened nonnuclear attacks by Russia against the United States could generate crisis instability, most likely by incidental strikes against dual-use U.S. communication capabilities. (In the more distant future, if China develops significant counterforce capabilities, it too could generate crisis instability through such attacks, though that possibility is not considered further here.)

The United States has acknowledged three "layers" of capabilities for sending employment orders to deployed nuclear forces: satellites, ground-based transmitters, and airborne transmitters.[79] The two U.S. satellite constellations for communicating with nuclear forces—the legacy Milstar system and newer AEHF system—are dual use. Because these satellites are in high-altitude geostationary orbits, there would be particular challenges in attacking them (including the possibility of evasive maneuvering by the target satellite in the time required for a direct-ascent weapon to reach it after launch). Such challenges notwithstanding, these satellites are likely to be vulnerable soon—if they are not already.[80] Russia has reportedly preserved—and may be enhancing—legacy Soviet direct-ascent ASAT weapons able to reach geostationary orbit and, in 2015, demonstrated an apparent co-orbital capability against satellites in that orbit.[81]

The United States also operates two networks of dual-use ground-based

77. "China Plans to Launch Test Satellite for Missile Defense," Japan Economic Newswire, August 24, 2015.
78. Indeed, media reports claim that both China's and Russia's early-warning satellites have the capability to contribute to missile-defense operations. See ibid.; and Graham, "Soyuz 2-1B Launches Tundra Missile Detection Spacecraft."
79. It is possible that the United States has additional classified systems. Because only a handful of technologies can communicate over long distances, however, any such systems would be likely to suffer from vulnerabilities similar to those of acknowledged systems.
80. The U.S. intelligence community assesses that "Russian and Chinese destructive ASAT weapons probably will reach initial operational capability in the next few years." This wording may suggest that nondestructive ASAT weapons may already be operational. See Coates, "Worldwide Threat Assessment of the U.S. Intelligence Community," p. 13.
81. Brian Weeden, "Dancing in the Dark Redux: Recent Russian Rendezvous and Proximity Operations in Space," *Space Review*, October 5, 2015, http://www.thespacereview.com/article/2839/1; and Arbatov, Dvorkin, and Topychkanov, "Entanglement as a New Security Threat," pp. 33–35.

transmitters that it can use to send employment orders to nuclear forces. The Fixed Submarine Broadcast System appears to comprise nine transmitters located mostly around the peripheries of the Atlantic and Pacific Oceans.[82] The High Frequency Global Communications System for communicating with bombers (and perhaps other nuclear delivery systems, too) consists of thirteen transmitters spread across the globe.[83] All of these transmitters are large fixed structures that (with one exception) are located near coasts, making them vulnerable to Russian sea- and air-launched cruise missiles, in particular.[84]

In a conventional conflict against NATO, Moscow might attack U.S. communication assets in an effort to further its warfighting goals. Russian strategists "can hardly imagine [such a] conflict failing to spread from the Euro-Atlantic region to the Far East-Pacific."[85] As a result, even in a European conflict, Russia might not limit its attacks to U.S. communication assets located in or around Europe. In fact, Russia could plausibly launch incidental strikes (most likely in a series of waves) against dual-use land-based transmitters spread around the Euro-Atlantic and Asia-Pacific areas, and, even more significantly, against perhaps three out of the four AEHF satellites (depending on exactly how the constellation is configured after Milstar satellites are retired). Washington would surely have little confidence that any remaining space- and land-based assets for communicating with nuclear forces would survive for long.

In this scenario, the United States would become critically dependent on E-4B and E-6B aircraft, which are designed to protect national and military leaders and facilitate communications with both nuclear and nonnuclear forces.[86] Indeed, a recent U.S. Strategic Command exercise, Global Thunder 2018, involved an adversary's attacking U.S. nuclear C3I assets until "the last thing remaining [was] the jet."[87] For now, these aircraft would likely be surviv-

82. U.S. Department of the Navy, "Submarine Communications Master Plan."
83. Dwayne Harris, "HFGCS Status" (Boston: Rockwell Collins, February 4, 2010), p. 5, http://www.hfindustry.com/meetings_presentations/presentation_materials/2010_feb_hfia/presentations/HFGCS_HFIA_Feb_2010.pdf.
84. The exception is a High Frequency Global Communications System transmitter in Nebraska. Given its location, however, it is unlikely to be involved in directing operations involving forward-deployed aircraft.
85. Alexei Arbatov, "Gambit or Endgame? The New State of Arms Control" (Moscow: Carnegie Moscow Center, March 2011), p. 6, http://carnegieendowment.org/files/gambit_endgame.pdf.
86. Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, "Nuclear Matters Handbook 2016" (Washington, D.C.: Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, 2016), p. 75, https://www.acq.osd.mil/ncbdp/nm/NMHB/docs/NMHB_2016-optimized.pdf.
87. Quoted in Sydney J. Freedberg Jr., "When the Football Comes Out, Who Watches the Presi-

able, because they could probably be protected by friendly forces while operating from U.S. airspace.

The survivability prospects of E-4B and E-6B aircraft over the longer term, however, are questionable. Because these aircraft use modified commercial airframes, they lack both the speed to escape threats and the stealth characteristics to avoid detection. Indeed, given that their fundamental purpose is communications, their eventual replacements could not be stealthy either. Russia, therefore, may be able to develop capabilities, such as long-range air-to-air weapons, that could threaten communication aircraft, even while operating over the United States. If so, incidental attacks on these aircraft—or even, perhaps, apparent preparations for such attacks—could generate crisis instability by appearing to be an attempt to undermine the U.S. nuclear deterrent by cutting off the ability of national leadership to communicate with deployed nuclear forces.

ESCALATION REDUX

Misinterpreted warning, the damage-limitation window, and crisis instability are not mutually exclusive. Multiple escalation pressures could arise simultaneously and even interact with one another. That said, for escalation to occur along any pathway, specific technological and doctrinal conditions would have to be fulfilled, as summarized in table 1. In abstract terms, each mechanism involves an "attacker" that launches or threatens nonnuclear attacks against a "target." Some conditions are necessary for the target to experience pressures to escalate the conflict. Others are contributory in that they increase the likelihood of escalation, but escalation can occur even if they are not fulfilled. For example, the target must have dual-use C3I capabilities, and those capabilities must be attacked or threatened for misinterpreted warning to occur. If the attacker has a counterforce nuclear doctrine (as Russia does), escalation is more likely. Nonetheless, escalation can still occur if the attacker does not plan for counterforce operations (as in the case of China).

Early Warning: Technical Vulnerabilities and Their Consequences

Two questions arise when assessing the severity of the escalation risks described above. First, how important to nonnuclear warfighting are the assets involved in nuclear C3I? The more important they are, the more likely they

---

dent?" *Breaking Defense*, November 9, 2017, https://breakingdefense.com/2017/11/stratcom-wargames-its-own-death-who-watches-the-president/.

Table 1. Technological and Doctrinal Conditions for Escalation to Occur as a Result of Entanglement

	Misinterpreted Warning	Damage-Limitation Window	Crisis Instability
Target's nuclear forces have been attacked by—or are perceived to be threatened by—attacker's nonnuclear weapons			××[a]
Target's nuclear C3I (command, control, communication, and intelligence) capabilities have been attacked by—or are perceived to be threatened by—attacker's nonnuclear weapons	××	××	××[a]
Target's nuclear delivery systems are dual use or superficially similar to nonnuclear delivery systems			×
Target's nuclear C3I capabilities are dual use	××	××	×
Attacker's nuclear doctrine calls for damage limitation	×		×
Target's nuclear doctrine calls for damage limitation	×	××	
Attacker's conventional warfighting doctrine calls for attacks against C3I capabilities	×	×	×

×× = necessary condition

× = exacerbating condition

[a]At least one of these conditions is necessary for crisis instability.

might be threatened or attacked in a conventional conflict. Second, how badly would strikes against dual-use enabling capabilities degrade the target's ability to prosecute a nuclear war? If the target's nuclear C3I system were highly resilient and limited strikes would do little to undermine its overall effectiveness, then the escalation risks of incidental strikes would probably be small. By contrast, if the loss of a few key enabling assets—in the worst case, just one—severely undermined the target's ability to conduct nuclear operations, escalation would be more likely.

This section demonstrates that the United States' early-warning system is deeply integrated into its conventional operations and that even limited strikes could lower its effectiveness significantly and so create serious escalation risks. Separately, it considers the risks of cyber interference with dual-use Chinese,

Russian, and U.S. early-warning capabilities. These risks have some important differences from those that might result from kinetic strikes.

Threats to early-warning assets are important in generating escalation risks through crisis instability, misinterpreted warning, and in the cases of Russia and the United States, the damage-limitation window. For reasons of space, threats to other enabling capabilities are not considered here, though are potentially no less significant. Indeed, in a real conflict, it is possible that multiple enabling systems could be attacked or threatened—potentially very early in a conflict, especially where ISR is concerned—magnifying escalation risks.

As with Russia, early warning would be necessary for the United States to execute any of the launch-under-attack options included in its nuclear war plans.[88] Under its policy of "dual phenomenology," Washington requires "two independent information sources using different physical principles" in assessing a potential attack.[89] To this end, the United States has deployed two distinct missile early-warning capabilities.[90] Space-based infrared detectors can identify the hot gases that are expelled from a ballistic missile while its motor is firing. Later in flight, large land-based radars can monitor the incoming reentry vehicle, potentially from a distance of thousands of kilometers.

If U.S. launch-under-attack plans include the option to launch nuclear weapons before any nuclear detonations on American soil—as was the case toward the end of the Cold War and appears to be true today—then the United States' early-warning architecture has no redundancy at the systems level; the loss of early-warning data from either satellites or radars could prevent Washington from meeting its own requirement for dual phenomenology.[91]

Threats to U.S. space-based early-warning assets

In 2018, the United States completed deployment of the Space-Based Infrared System (SBIRS) to replace the legacy Defense Support Program system for space-based early warning. The SBIRS constellation comprises six satel-

88. Bureau of Arms Control, Verification, and Compliance, "U.S. Nuclear Force Posture and De-Alerting," fact sheet (Washington, D.C.: U.S. Department of State, December 14, 2015), https://web.archive.org/web/20170101112527/https://www.state.gov/t/avc/rls/250644.htm.
89. Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, "Nuclear Matters Handbook 2016," p. 76.
90. In addition, various systems can detect the detonations of nuclear warheads, but these are less useful for enhancing force survivability.
91. According to the U.S. State Department, "The President would have less than 30 minutes in which to make a decision to launch our ICBMs under attack." This timeline strongly suggests that a launch could take place before incoming warheads detonated. Bureau of Arms Control, Verification, and Compliance, "U.S. Nuclear Force Posture and De-Alerting." On Cold War policy, see Blair, *The Logic of Accidental Nuclear War*, pp. 168, 192.

lites.[92] Four dedicated SBIRS GEO satellites are in geostationary orbits, about 36,000 kilometers above fixed points near the Equator. In addition, to provide coverage of the northern polar region, two more SBIRS HEO detectors are hosted by "classified" satellites, whose primary purpose is reportedly electronic-intelligence collection, in highly elliptical orbits.[93] These satellites spend most of their orbits in the Northern Hemisphere, reaching latitudes as high as 65°N.

As with U.S. communication satellites, it is likely that if SBIRS satellites are not already vulnerable, they will be soon.[94] The U.S. intelligence community assesses that both China and Russia are "advancing directed-energy weapons technologies for the purpose of fielding ASAT weapons that could blind or damage sensitive space-based optical sensors, such as those used for . . . missile defense."[95] Moreover, like Russia, China is funding the development of direct-ascent ASAT weapons and, in 2013, probably tested an ASAT weapon that may be capable of threating geostationary satellites.[96]

In a conventional conflict, an adversary could have at least two significant motivations for launching incidental attacks against the United States' SBIRS constellation. First, the electronic-intelligence collection satellites that reportedly host SBIRS HEO detectors are in orbits ideally suited for monitoring military activities in Russia's north, making them potential targets. One particularly strong motivation for Moscow to attack them might be to interfere with U.S. efforts to collect intelligence on the movements of the surface ships and submarines of Russia's Northern Fleet, which is based inside the Arctic Circle. In such strikes, the SBIRS HEO detectors would be collateral damage.

Second, China or Russia could target the SBIRS constellation because of its role in enabling nonnuclear operations. The constellation's most important

92. The United States has purchased additional satellites for replenishment purposes; more than six satellites, therefore, may temporarily be in orbit.
93. Office of the Secretary of Defense, "Report to the Defense and Intelligence Committees of the Congress of the United States on the Status of the Space Based Infrared System Program" (Washington, D.C.: U.S. Department of Defense, March 2005), p. 31, http://nsarchive.gwu.edu/NSAEBB/NSAEBB235/42.pdf; and Michel Capderou, *Handbook of Satellite Orbits: From Kepler to GPS*, trans. Stephen Lyle (Cham, Switzerland: Springer, 2014), p. 428 n. 133.
94. Attacks against ground-based uplinks or downlinks are also a threat, but are not considered further here.
95. Coates, "Worldwide Threat Assessment of the U.S. Intelligence Community," p. 13.
96. Brian Weeden, "Through a Glass, Darkly: Chinese, American, and Russian Anti-Satellite Testing in Space" (Broomfield, Colo.: Secure World Foundation, March 17, 2014), pp. 4–19, https://swfound.org/media/167224/through_a_glass_darkly_march2014.pdf; and U.S.-China Economic and Security Review Commission, "2015 Report to Congress" (Washington, D.C.: U.S. Government Printing Office, November 2015), pp. 292–298, https://www.uscc.gov/sites/default/files/annual_reports/2015%20Annual%20Report%20to%20Congress.PDF.

such functions are providing early warning of, and cueing defenses against, nonnuclear ballistic missiles. In general, the more satellites were attacked, the more the performance of U.S. defenses would be degraded. SBIRS satellites are involved in other nonnuclear missions, including "intelligence collection" and "battlespace characterization," which includes "battle damage assessment, suppression of enemy air defense, [and] enemy aircraft surveillance."[97] In a few circumstances, these auxiliary functions could be sufficiently important to motivate an adversary to launch incidental attacks. For example, China might attack SBIRS satellites because of their ability to detect nonnuclear ballistic missiles early in flight and hence provide targeting data that the United States would find useful if it sought to hunt the mobile launchers from which such missiles were being launched.[98]

Not only might China or Russia attack SBIRS satellites in a conventional conflict, but such attacks—even if limited—could have serious negative implications for the United States' ability to monitor launches of the adversary's nuclear-armed ballistic missiles.[99] With six satellites, the SBIRS constellation can be—and, after the retirement of the remaining Defense Support Program satellites, presumably will be—configured so that most areas from which nuclear-armed missiles might plausibly be launched are monitored by at least three or four satellites at all times, providing some margin of redundancy. In practice, however, this margin could be worn away quickly. If Beijing or Moscow sought, in a conventional conflict, to undermine U.S. missile defenses by degrading the SBIRS constellation to point where it could not monitor nonnuclear missile launches from, respectively, Eastern China or Western Russia, the United States would also lose the capability to monitor the majority of its adversary's nuclear forces continuously from space.

The margin of redundancy for some potential launch sites is even thinner. For example, if Russia destroyed just two SBIRS satellites—either of the host satellites for SBIRS HEO detectors, and the western-most SBIRS GEO satellite (which would contribute significantly to ballistic missile defense operations in Europe)—it would deprive the United States of the space-based capability

97. Office of the Secretary of Defense, "Report to the Defense and Intelligence Committees of the Congress of the United States on the Status of the Space Based Infrared System Program," p. 4.
98. Morgan, "Deterrence and First-Strike Stability in Space," p. 20.
99. This discussion is based on the author's own analysis using NASA's General Mission Analysis Tool orbital modeling software and data about satellite orbits from Chris Peat, *Heavens Above* (website), http://www.heavens-above.com; and Jonathan McDowell, "Geostationary Orbit Catalog," *Jonathan's Space Report*, n.d., http://www.planet4589.org/space/log/geo.log. It assumes that, after legacy Defense Support Program satellites have been retired, SBIRS GEO 4 will be placed in an orbit at or near 66°E, where a Defense Support Program satellite is currently located.

to continuously monitor potential Russian SSBN patrol areas in the North Atlantic Ocean close to Europe.

Moreover, the SBIRS constellation features a single-point vulnerability: the United States could not continuously monitor the northern polar region from space if either of the SBIRS HEO detectors were rendered inoperable. With just one of these detectors in operation, there would be slightly more than four-and-a-half hours each day during which the United States had no coverage of the northern polar region or only partial coverage. Gen. William Shelton, then commander of U.S. Space Command, was almost certainly referring to this weakness when, in 2014, he acknowledged, without further explanation, the existence of a single-point vulnerability in the SBIRS constellation.[100]

Historically, monitoring the northern polar region has not been a U.S. priority, presumably because so much of it used to be covered by ice year round that it was an undesirable area from which to launch ballistic missiles.[101] Indeed, until the first SBIRS HEO detector was launched in 2006, the United States relied solely on land-based radars for this task. As climate change further reduces the sea ice coverage of the Arctic Ocean, however, especially during summer, monitoring the northern polar region is probably becoming more important.

THREATS TO U.S. LAND-BASED EARLY-WARNING ASSETS

The United States operates six land-based early-warning radars designed primarily to detect missile attacks against the United States: five PAVE PAWS radars are located in California, Massachusetts, Greenland, the United Kingdom, and Alaska, where a COBRA DANE radar is also based.[102] All of these ballistic missile early-warning radars are large and immobile, and hence potentially vulnerable to precise conventional weapons, including air- and sea-launched cruise missiles. In general, opening a complete hole in the U.S. network of BMEWRs would require the destruction of at least two or three radars.[103]

---

100. William L. Shelton, "Space and Cyberspace—Foundational Capabilities for the Joint Warfighter and the Nation," speech at the Air Force Association Air Warfare Symposium, Orlando, Florida, February 21, 2014, http://web.archive.org/web/20141225171206/http://www.afspc.af.mil/library/speeches/speech.asp?id=747.

101. Russian SSBNs reportedly had some capability to launch missiles from under relatively thin ice. Valery E. Yarynich, "C^3: Nuclear Command, Control Cooperation" (Washington, D.C.: Center for Defense Information, May 2003), p. 147, https://www.scribd.com/doc/282622838/C3-Nuclear-Command-Control-Cooperation.

102. Technically, after PAVE PAWS radars are upgraded to contribute to missile-defense operations, they are renamed Upgraded Early Warning Radars. Forward-deployed missile-defense radars, such as the AN/TPY-2, could also contribute to early warning.

103. In theory, the destruction of the radar in either California or Massachusetts would create such

Although the primary mission of U.S. BMEWRs is to detect and track an incoming nuclear strike, they also contribute significantly to two nonnuclear operations. First, they have a significant role in tracking space objects, including U.S. satellites and potentially Chinese and Russian ASAT weapons. As a result, Beijing or Moscow might plausibly attack U.S. BMEWRs to maximize both the effectiveness and consequences of ASAT operations.

Second, like early-warning satellites, the United States' BMEWRs have (or, in some cases, are currently being upgraded to gain) the capability to contribute to defending against nonnuclear ballistic missile strikes. Indeed, both China and Russia evince an interest in holding ground-based U.S. ballistic missile defense assets at risk.[104] That said, today at least, only one BMEWR—the one based at Fylingdales in the United Kingdom—could likely become involved in defending against Chinese or Russian nonnuclear ballistic missile strikes and so be subject to incidental attacks.[105] Given the location of other U.S. BMEWRs, the only Chinese or Russian ballistic missiles that they would be likely to track would be SLBMs or ICBMs, all of which are currently nuclear armed.

The U.S. Fylingdales BMEWR is not only the radar most likely to suffer an incidental attack; it is also the most important radar for providing early warning of a Russian nuclear strike. Because it is based so far east of the continental United States, this radar could detect Russian ICBM and SLBM launches from most deployment areas much earlier than other U.S. BMEWRs. As a result, especially if Russia succeeded in partially or completely disabling the SBIRS constellation, follow-on attacks against the Fylingdales radar could be particularly escalatory. Looking forward, if China or Russia eventually develops nonnuclear ICBMs or SLBMs, then U.S. BMEWRs other than Fylingdales could take on a significant role in nonnuclear missile defense operations, creating new targets for incidental attacks and thus potential triggers of escalation.

---

a hole, but it is currently unlikely that either of these radars and no other would be subject to incidental attacks. The discussion in this section is based on the author's own analysis using Google Earth. The author thanks Geoffrey Forden and Pavel Podvig for assistance with, respectively, visualizing ballistic missile trajectories and radar fans. Data about the capabilities of U.S. BMEWRs are available from Missile Defense Agency, "Elements: Sensors" (Washington, D.C.: U.S. Department of Defense, January 22, 2018), https://www.mda.mil/system/sensors.html.

104. Second Artillery Corps, "The Science of Second Artillery Campaigns," pp. 318, 396–397; and Andrew E. Kramer, "Russian General Makes Threat on Missile-Defense Sites," *New York Times*, May 3, 2012, http://www.nytimes.com/2012/05/04/world/europe/russian-general-threatens-pre-emptive-attacks-on-missile-defense-sites.html.

105. The author estimates that the Fylingdales radar, if angled at 3 degrees above the horizontal, could track an Iskander ballistic missile with a range of 500 kilometers fired from Kaliningrad to western Poland for about 150 kilometers of its trajectory.

Intriguingly, there may be an Asia-Pacific analogue to the Fylingdales radar—even if it is not a U.S. radar. In 2013, Taiwan commissioned a PAVE PAWS early-warning radar that it purchased from the United States. Taipei has stated that the sole purpose of this radar is to track Chinese short-range nonnuclear ballistic missiles.[106] This radar, however, does have an inherent capability to detect Chinese ICBMs early in flight. In fact, it could provide significantly more warning of a Chinese ICBM strike than any U.S. BMEWR, and a senior Taiwanese lawmaker has claimed that data from this radar is shared with the United States.[107] If this claim is correct, then incidental Chinese strikes against the radar could have significant escalation consequences. To be sure, China might attack this radar in the early phases of a conflict, while the war's outcome was still uncertain. At that juncture, the escalation risks would probably be modest, because China's incentives to use nuclear weapons would be minimal. If, however, China began to lose the conflict and subsequently attacked the United States' SBIRS satellites, then the already serious escalation consequences of attacking early-warning satellites would likely be compounded by China's earlier attack on the radar.

CYBER THREATS TO EARLY-WARNING SYSTEMS

There are credible reports of cyber interference with early-warning systems. Most notably, when Israel destroyed Syria's clandestine plutonium-production reactor in 2007, it reportedly first disabled the Syrian air-defense system using a variety of tools, including cyberweapons, to reduce risks to the aircraft conducting the attack.[108] To be sure, nuclear C3I networks in China, Russia, and the United States are presumably protected by much better cyber defenses than Syria's air-defense system was a decade ago. Nonetheless, these states have launched efforts to enhance the cyber defenses of networks used for nuclear C3I, implying that they believe cyber threats to them are credible; indeed, the United States military has said so explicitly.[109] Yet, eliminating cyber vul-

106. "Taiwan Deploys Advanced Early Warning Radar System," *Straits Times*, February 3, 2013, http://www.straitstimes.com/asia/taiwan-deploys-advanced-early-warning-radar-system.
107. "Long-Range Radar Budget Surges by NT$10 Billion," *China Post*, January 6, 2013, https://web.archive.org/web/20130108092745/http://www.chinapost.com.tw/taiwan/national/national-news/2013/01/06/366468/Long-range-radar.htm.
108. David A. Fulghum, Robert Wall, and Amy Butler, "Cyber-Combat's First Shot: Israel Shows Electronic Prowess: Attack on Syria Shows Israel Is Master of the High-Tech Battle," *Aviation Week & Space Technology*, November 26, 2007, pp. 28–31. See also Joshua Berlinger and Juliet Perry, "China Tried to Hack Group Linked to Controversial Missile Defense System, U.S. Cybersecurity Firm Says," *CNN*, April 27, 2017, http://www.cnn.com/2017/04/27/asia/china-south-korea-thaad-hack/.
109. See, for example, Michael Pillsbury, "The Sixteen Fears: China's Strategic Psychology," *Sur-*

nerabilities entirely may be impossible. The U.S. Defense Science Board, for example, has stated baldy that it is "impossible" for the U.S. Department of Defense to fully defend its networks.[110]

The existing literature on cyber threats to early-warning systems has not considered the possibility that dual-use early-warning capabilities might be subject to incidental cyber interference for the purpose of influencing the outcome of a conventional war.[111] (Because some physical early-warning assets in China, Russia, and the United States are dual use, at least some of the networks that support them must also be dual use.[112]) The term "cyber interference" is used here to include both cyber espionage (gathering information for intelligence purposes without damaging the operation of the target system) and cyberattack (attempting to undermine the target system's functionality by compromising the integrity or availability of its data).

The severity of the escalation risks stemming from incidental cyber interference with early-warning capabilities depends on at least two factors. One factor, as Erik Gatrzke and Jon Lindsay note, is whether the target detects the cyber interference.[113] The other is whether, if the interference is detected, the target correctly assesses the attacker's intent.

---

vival, Vol. 54, No. 5 (October/November 2012), p. 157, doi:10.1080/00396338.2012.728351; "Cyber Security Units to Protect Russia's Nuclear Weapons Stockpiles," *RT*, October 17, 2014, https://www.rt.com/news/196720-russia-missile-forces-cybersecurity/; and Benjamin D. Katz, "U.S. Beefs Up Cyber Defenses to Thwart Hacks of Nuclear Arsenal," *Bloomberg*, March 24, 2016, https://www.bloomberg.com/news/articles/2016-03-24/u-s-beefs-up-cyber-defenses-to-thwart-hacks-of-nuclear-arsenal.

110. Defense Science Board, U.S. Department of Defense, "Task Force Report: Resilient Military Systems and the Advanced Cyber Threat" (Washington, D.C.: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, January 2013), p. 6, https://www.acq.osd.mil/dsb/reports/2010s/ResilientMilitarySystemsCyberThreat.pdf.

111. This literature focuses on two deliberate escalation risks. First, a malicious third party could try to catalyze a nuclear war between two nuclear-armed states by generating false warning of an incoming nuclear attack. Second, a state planning to launch a nuclear attack might first "blind" its opponent's early-warning system. See Global Zero Commission on Nuclear Risk Reduction, "De-Alerting and Stabilizing the World's Nuclear Force Postures" (Washington, D.C.: Global Zero, 2015), p. 30, http://www.globalzero.org/files/global_zero_commission_on_nuclear_risk_reduction_report_0.pdf; Andrew Futter, *Cyber Threats and Nuclear Weapons: New Questions for Command and Control, Security, and Strategy* (London: Royal United Services Institute for Defence and Security Studies, July 2016), pp. 24–25, https://rusi.org/sites/default/files/cyber_threats_and_nuclear_combined.1.pdf; and Stephen J. Cimbala, "Nuclear Cyberwar and Crisis Management," *Comparative Strategy*, Vol. 35, No. 2 (2016), p. 119, doi:10.1080/01495933.2016.1176458.

112. At the very least, the networks used to determine whether an incoming weapon was conventional or nuclear (and at all prior stages of the early-warning process) must be dual use.

113. Erik Gartzke and Jon R. Lindsay focus on deliberate attacks intended to undermine the target's nuclear deterrent. The escalation dynamics resulting from those attacks and incidental attacks would be different. See Gartzke and Lindsay, "Thermonuclear Cyber War," *Journal of Cyber Security*, Vol. 3, No. 1 (March 2017), pp. 37–48, doi:10.1093/cybsec/tyw017.

In even a modest conventional conflict, a state's temptation to conduct cyber espionage against an enemy's C3I system could be very strong. In the case of dual-use early-warning networks, the state might focus on detecting its opponent's potential weaknesses—such as radars that were inoperative or performing poorly—so the state could exploit them to enable more effective offensive operations. Such cyber espionage could have escalation consequences only if the target discovered it. In this case, the espionage could contribute to misinterpreted warning, because the target might believe that its opponent was looking for weaknesses prior to using nuclear weapons. The exact consequences, though, would presumably depend on what the target believed the cyber espionage had revealed. For example, if Russia believed that the United States had discovered a serious weakness in its early-warning system, Moscow's confidence in the survivability of its nuclear forces could diminish, generating crisis instability on top of misinterpreted warning. By contrast, if Russia believed that the United States had failed to acquire anything of significance, the escalation consequences might be much more modest.

Cyberattacks designed to facilitate nonnuclear strikes by undermining the operation of an adversary's early-warning capabilities could also precipitate escalation. Once again, the attack could prove escalatory only if the target detected it. If a state did conclude that its early-warning system had been subject to a cyberattack, the escalation consequences could be as serious as if the system had been physically attacked, especially if the target believed that the damage could not be reversed quickly. In fact, the consequences might even be more serious because a cyberattack against a critical network (one responsible for fusing data from multiple sources, say) could disable an entire early-warning system, whereas kinetic strikes would have to pick off sensors one by one.

The risk of escalation could be further exacerbated by the challenges facing the target in determining the attacker's intent. Fully understanding the purpose of complex malware can be difficult and time consuming, and the target might be uncertain about its capabilities for a significant length of time—allowing considerable scope for worst-case thinking. For example, even if the malware were capable only of espionage, the target might worry it also contained a "kill switch" able to disable an early-warning system after activation. To create yet more uncertainty, a single penetration into a network can be used to insert multiple "payloads." For this reason, even if the target believed that ongoing interference was limited to espionage, it might worry that the vulnerability used by the attacker could be exploited for more nefarious ends (at least until that vulnerability had been identified and fixed).

In the final analysis, there would be at least two important differences between the escalation risks resulting from cyber interference with and physical attacks on dual-use early-warning systems. First, a physical attack on an early-warning asset would be significantly more difficult to conceal than cyber interference (even if not all physical attacks are equally obvious). Unlike plausible physical attacks, therefore, cyberattacks on early-warning systems might go undetected and have no escalation consequences. Second, with physical attacks on early-warning assets, the risk of inadvertent escalation would stem from the dual-use nature of the target. With cyber interference, this ambiguity would still exist but would be compounded by possible uncertainty about the interference's purpose. This "double ambiguity" is a major reason why the escalation risks of U.S. nonnuclear operations against China or Russia would be greater than previous academic analyses have suggested. It means that even limited cyber espionage, if detected, could prove highly escalatory.

Policy Implications

In spite of the magnitude of the dangers, risk reduction is likely to prove extremely challenging. China, Russia, and the United States would be unlikely to agree to meaningful limits on nonnuclear capabilities designed to threaten potential adversaries' C3I assets because each state views such capabilities as critical for both conventional warfighting and deterrence. Moreover, each state is—or may become—resistant to disentangling its nuclear and nonnuclear forces and C3I assets. Russia's objection, according to Alexey Arbatov, is simply the financial costs of separation.[114] Some Chinese scholars, meanwhile, have argued that separating nuclear and nonnuclear forces and C3I assets could make U.S. attacks against Chinese nonnuclear capabilities less risky and hence more likely (even if these scholars also argue that China's adoption of dual-use capabilities was originally motivated by convenience and not strategy).[115] Indeed, the same logic may even end up holding sway in Washington. There is no evidence that the United States' use of dual-use C3I assets (or dual-use aircraft, for that matter) was motivated by anything other than convenience and cost. If, however, there was ever a serious discussion about

114. Alexey Arbatov, "Non-Nuclear Weapons and the Risk of Nuclear War: A Russian Perspective," discussion at the Carnegie Endowment for International Peace, Washington, D.C., November 29, 2017, http://carnegieendowment.org/2017/11/29/non-nuclear-weapons-and-risk-of-nuclear-war-russian-perspective-event-5762 (in particular, the comments at 38:44 of the recording).
115. Zhao and Li, "The Underappreciated Risks of Entanglement," p. 68.

separating nuclear and nonnuclear C3I then it is not difficult to imagine advocacy for entanglement on deterrence grounds.

Yet, Beijing, Moscow, and Washington should still confront the question of whether the advantages of entanglement—both financial and strategic—are worth the escalation risks. After all, if the escalation risks are too great, then any benefits will be outweighed by an increase in the likelihood and probable costs of a war. If this article is correct—if the escalation risks are greater than widely realized and likely to increase further—then China, Russia, and the United States may already be on the wrong side of the line.

UNDERSTANDING AND RAISING AWARNESS OF THE RISKS

A first-order task for Washington, Beijing, and Moscow, therefore, is to conduct their own analyses, most likely on a classified basis, of the potential benefits and risks of entanglement. These efforts should be informed by intelligence assessments about the extent to which potential adversaries' nuclear and nonnuclear forces and C3I assets are entangled, and about those rivals' perceptions of the intentions and capabilities of the state conducting the analysis. If such analyses concluded that the risks of entanglement did indeed outweigh the benefits, they could catalyze and inform the development of a risk-reduction strategy.

In principle, there are both unilateral and cooperative approaches to risk mitigation. Given the poor state of political relations between Washington and Beijing, and between Washington and Moscow, unilateral measures currently represent the only feasible starting point. Such measures certainly cannot eliminate the escalation risks of entanglement, but they could help to mitigate them and slow their rate of increase.

In this vein, the simplest risk-reduction measure would be to raise awareness, within governments and militaries, of the challenges created by entanglement for assessing an adversary's intent and, importantly, for the adversary in assessing the state's own intent. Given that crisis instability and misinterpreted warning are mediated by perceptions—or rather misperceptions—about the intent behind incidental strikes or threats, drawing the attention of decisionmakers to the difficulties of assessing intent might encourage restraint in a conflict and so help counteract inadvertent escalation pressures. Greater awareness of the risks could also catalyze peacetime preparations, such as enhancing the survivability of C3I assets, that might reduce the dangers associated with incidental strikes should a war occur. Such preparations might simultaneously mitigate the escalation risks resulting from the existence of

the damage-limitation window (which are not driven by misjudgments about intent).

To this end, China, Russia, and the United States could set up risk-reduction teams within their defense establishments.[116] Most important, during crises or conflicts, these teams could advise national and military leaders on the risks associated with entanglement and on ways to manage them. In peacetime, they could be tasked with ensuring that escalation risks were factored into both war planning and acquisition decisions for new strategic weapons and C3I capabilities (the teams could, for example, assess the different alternatives under consideration for their escalation implications, and be entitled to propose other options or object to the program entirely).

Ultimately, of course, high-level civilian or military leaders would be responsible for making decisions after considering escalation risks alongside more traditional strategic, military, and financial considerations. Risk-reduction teams, therefore, would have to be bureaucratically empowered (by being led by a suitably senior official, for example) to ensure their advice was heard. Such teams would also benefit by being made up from a broad range of experts, including civilian strategists, military planners, and intelligence officials with deep knowledge of potential adversaries' thinking.

In addition to their other tasks, risk-reduction teams could be tasked, in peacetime, with proposing unilateral risk-reduction measures. Changes to declaratory policy (which could be accomplished rapidly) and C3I system design (which could take years to implement) are examples of two different but complementary approaches.

DECLARATORY POLICY

Declaratory policy is one tool for deterring incidental attacks on C3I assets by underscoring the risks. It is possible that the civilian officials or military officers responsible for authorizing such attacks might not appreciate the potential for their intentions to be misinterpreted. Such officials could hold very senior positions (kinetic ASAT attacks, in particular, might require authoriza-

116. James M. Acton, "Technology, Doctrine, and the Risk of Nuclear War," in Nina Tannenwald, Acton, and Jane Vaynman, *Meeting the Challenges of the New Nuclear Age: Emerging Risks and Declining Norms in the Age of Technological Innovation and Changing Nuclear Doctrines* (Cambridge, Mass.: American Academy of Arts and Sciences, 2018), pp. 54–55, https://www.amacad.org/multimedia/pdfs/publications/researchpapersmonographs/New-Nuclear-Age_Emerging-Risks/New-Nuclear-Age_Emerging-Risks.pdf. This idea was inspired by Posen, *Inadvertent Escalation*, pp. 212–218.

tion from a head of state) and might not know that such assets were typically dual use; even if they did, they might not appreciate the implications.

The 2018 U.S. Nuclear Posture Review's threat to use nuclear weapons in response to attacks on nuclear C3I assets is presumably an attempt to warn potential adversaries about these implications. The disproportionate nature of this threat, however, risks its being dismissed by Beijing and Moscow as bluster. Instead, a somewhat vaguer formulation might ultimately prove more effective. For example, Washington could state that it considers dual-use communication and early-warning assets an integral part of its nuclear C3I system and would respond to attacks on them accordingly (Beijing and Moscow could make similar statements). As with all declaratory policy, such statements might influence potential adversaries' thinking more effectively if they were repeated periodically by very senior officials.

TOWARD A MORE RESILIENT C3I ARCHITECTURE

Over the longer term, states could also develop C3I architectures that were both less likely to be subject to incidental attacks and more survivable if they were. Some analysts have suggested creating at least two separate C3I systems—one for nuclear or "strategic" operations and one (or more) for all other operations.[117] Even putting the costs of this idea aside, such disaggregation would reduce risks only if Washington, say, could convince Beijing and Moscow that it had separated nuclear and nonnuclear C3I functions, which would be no easy task. If the United States failed to do so, disaggregation could increase risks because the escalation consequences of China's or Russia's attacking C3I assets that were involved only in nuclear operations—out of the incorrect belief that they also enabled conventional operations—could be more severe than the consequences of attacking dual-use assets.

A somewhat different approach for early warning would be to create space-based capabilities that were less likely to be subject to incidental attack because they were incapable of contributing significantly to any mission other than detecting the launch of an adversary's missiles (whether nuclear or nonnuclear). In particular, as a matter of basic optics, physically small infrared

117. Elbridge Colby, "From Sanctuary to Battlefield: A Framework for a U.S. Defense and Deterrence Strategy for Space" (Washington, D.C.: Center for a New American Security, January 2016), p. 22, https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Space-Report_16107.pdf; and Todd Harrison, "The Future of MILSATCOM" (Washington, D.C.: Center for Strategic and Budgetary Assessments, 2013), pp. 40–42, http://csbaonline.org/uploads/documents/Future-of-MILSATCOM-web.pdf.

detectors would be incapable of producing the kind of high-resolution imagery that would be most useful for cueing missile defenses and detecting the exact location of mobile missile launchers.[118] Because this limitation was the result of an observable and immutable property of the hardware, Washington, say, might be able to persuade Beijing and Moscow that it was real.

Another key advantage of small detectors is that they would not require their own satellite buses (which are generally very expensive to design and manufacture), but could instead be hosted by satellites used for other purposes. In this way, it might be possible to deploy them affordably in large numbers—tens, perhaps—creating a resilient architecture that would be the early-warning equivalent to AFSATCOM.[119] Although this author's judgment is that this kind of "dispersed" early-warning system would reduce the risks associated with incidental attacks, important challenges and trade-offs that deserve further study would arise.

For example, if the host satellites were attacked to undermine their primary function, their associated early-warning detectors would almost inevitably also be destroyed. To be sure, the likelihood of such attacks could be reduced by choosing host satellites that might not otherwise be targets (such as weather or commercial noncommunication satellites), and the consequences of such attacks would be mitigated by having multiple detectors in orbit. Nonetheless, a dispersed system could not eliminate the risks associated with incidental attacks.

Separately, deploying a dispersed system in addition to more capable dedicated early-warning satellites, such as SBIRS, might increase an adversary's incentives to attack the dedicated satellites (by reducing the escalation risks of doing so), and would be more expensive than fielding either system alone.[120] By contrast, deploying a dispersed system instead of dedicated satellites would lower the effectiveness of missile defenses.

Reducing an adversary's incentives to launch incidental attacks against space-based communication assets would be more difficult. Although a system that was capable of transmitting data only at low rates would be somewhat more useful for nuclear than nonnuclear operations, there would be no obvious way of demonstrating to adversaries that such a limitation was real

118. Eugene Hecht, *Optics*, 5th ed. (Boston: Pearson, 2017), p. 493.
119. Acton, "Command and Control in the Nuclear Posture Review."
120. If a dispersed system could be kept secret, its existence could not incentivize attacks against dedicated satellites. Under this approach, the United States would obviously not attempt to convince potential adversaries that the dispersed system was ineffective for any mission other than detecting missile launches.

and permanent. Instead, risk-reduction efforts could focus on mitigating the consequences of incidental attacks against space-based communication assets by enhancing their resilience. One approach would be to create an upgraded version of AFSATCOM by hosting small communication transponders for nuclear operations on tens of satellites used for other purposes (though, again, trade-offs similar to those associated with a dispersed early-warning system would arise).

Conclusion

As U.S.-Chinese and U.S.-Russian tensions have increased, albeit nonmonotonically, since the mid-2000s, warnings about the escalation risks that are inherent to the way that the United States would likely approach a great-power conflict have grown louder.[121] This focus on American doctrine and technology, however, has largely obscured another danger: the emerging Chinese and Russian ways of fighting wars are inherently escalatory too.

Both China and Russia, like the United States, seek to threaten potential adversaries' C3I assets and are improving their capabilities to do so. Because many enabling assets are dual use, however, attacks against them could, in the event of a conflict, degrade the target's nuclear C3I system just as a nuclear war was becoming all too imaginable. Crisis instability is one potential consequence. Indeed, its risks are more serious than generally understood because C3I assets that are space-based or distant from potential theaters of conflict could be subject to incidental kinetic attack or cyber interference. Additionally, C3I vulnerability could generate two other escalation pressures—misinterpreted warning and the damage-limitation window—that have not been previously discussed. Attacks against ISR assets, which would be likely in a major conflict, would exacerbate the risks by complicating the task of assessing an attacker's intent and by raising concerns about follow-on attacks against dual-use early-warning and communication assets.

In the future, the extent of entanglement—and hence the magnitude of these escalation risks—is likely to increase. Early-warning capabilities are likely to become more entangled with nonnuclear weapons as China and Russia modernize early-warning systems, and especially if one of them or the United States deploys nonnuclear SLBMs, ICBMs, or long-range hypersonic boost-

121. This literature has a broader focus than the vulnerability of nuclear forces and C3I assets. See, for example, Keir A. Lieber and Daryl G. Press, "The Nukes We Need: Preserving the American Deterrent," *Foreign Affairs*, Vol. 88, No. 6 (November/December 2009), p. 43.

glide weapons, which could be monitored in flight with capabilities primarily designed to detect a nuclear strike. Other nuclear C3I capabilities could also become more deeply integrated into nonnuclear missions. Because dual-use weapon-delivery systems may become more common, for example, the overlap between nuclear and nonnuclear enabling capabilities (such as communication and mission planning systems) is also likely to increase.

Nonnuclear threats to dual-use C3I capabilities are also likely to become more serious. For example, as part of U.S. efforts to enable forces to "take advantage of freedom of action in one domain to . . . challenge an adversary in another," the United States could develop or enhance ASAT capabilities—including kinetic ones, perhaps—for targeting dual-use Chinese and Russian communication and ISR satellites.[122] Meanwhile, if Beijing or Moscow develops long-range nonnuclear hypersonic boost-glide weapons, it may be able to threaten the uplinks and downlinks for U.S. satellites across the world, including in the continental United States—potentially endangering the functionality of multiple dual-use U.S. C3I systems.

If these risks are to be ameliorated—or, at the very least, if their rate of increase is to be stemmed—China, Russia, and the United States will first have to conclude that the risks of entanglement outweigh the benefits. If one or more of them reaches that conclusion then, for the time being, unilateral risk-reduction measures (including the use of declaratory policy to underscore the risks of attacking dual-use C3I assets and the development of more resilient C3I systems) offer the most promising way forward. Establishing risk-reduction teams would help institutionalize and inform these efforts and, perhaps most importantly, raise awareness of the risks within governments and militaries, thus helping to mitigate them.

Over the longer term, cooperative risk-reduction measures could be adopted to further mitigate the risks, particularly the threat to dual-use C3I capabilities. Although there is little prospect of such measures being negotiated today, the level of interest in them could rise in the future—because of a thaw in political relations, perhaps, or a dangerous crisis that shocked political leaders into action. States could, for example, commit not to engage in cyber interference with one another's nuclear C3I systems.[123] They could also agree to prohibit

122. Air-Sea Battle Office, "Air-Sea Battle," 4. See also Biddle and Oelrich, "Future Warfare in the Western Pacific," pp. 44–45; and Christensen, "The Meaning of the Nuclear Evolution," p. 472.
123. Richard J. Danzig, "Surviving on a Diet of Poisoned Fruit: Reducing the National Security Risks of America's Cyber Dependencies" (Washington, D.C.: Center for a New American Security, July 2014), pp. 24–27, https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Poisoned Fruit_Danzig.pdf.

the testing of ASAT weapons capable of threatening objects in geostationary orbit (where the most important space-based nuclear C3I assets are located).[124] Such prohibitions could prove effective if each participant assessed that the costs of violating the agreement—most obviously, the possibility that potential adversaries would engage in reciprocal violations—outweighed the disadvantages of compliance.

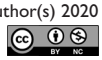
To make such prohibitions workable, significant technical challenges would need to be overcome. How would a prohibition against interfering with nuclear C3I systems be defined? What command-and-control systems would be covered given that so many of them are dual use? Similarly, what kind of weapons—precisely—would be included in a ban on testing ASAT weapons capable of reaching geostationary orbit? While challenging to answer, these questions are not necessarily unanswerable. In fact, the process of designing unilateral risk-reduction measures might stimulate and facilitate thinking about cooperative risk reduction by creating enhanced understanding of the risks associated with entanglement as well as the expertise to manage them. In this way, by embarking on unilateral risk-reduction processes now, China, Russia, and the United States could better position themselves to take advantage of any political opportunities for negotiations on cooperative measures that might arise in the future.

124. Arbatov, Dvorkin, and Topychkanov, "Entanglement as a New Security Threat," pp. 40–41.

Special issue on *Quantizing international relations*

SECURITY DIALOGUE

Security Dialogue
2020, Vol. 51(5) 450–466



DOI: 10.1177/0967010620901909
journals.sagepub.com/home/sdi


To 'see' is to break an entanglement: Quantum measurement, trauma and security

K M Fierke
University of St Andrews, UK

Nicola Mackay
Independent researcher, UK

Abstract

This article seeks to explore the quantum notion that to 'see' an entanglement is to break it in the context of an 'experiment' regarding the ongoing impact of traumatic political memory on the present. The analysis is a product of collaboration over the past four years between the two authors, one a scholar of international relations, the other a therapeutic practitioner with training in medical physics. Our focus is the conceptual claim that 'seeing' breaks an entanglement rather than the experiment itself. The first section explores a broad contrast between classical and quantum measurement, asking what this might mean at the macroscopic level. The second section categorizes Wendt's claim about language as a form of expressive measurement and explores the relationship to discourse analysis. The third section explores the broad contours of our experiment and the role of a somewhat different form of non-linear expressive measurement. In the final section, we elaborate the relationship between redemptive measurement and breaking an entanglement, which involves a form of 'seeing' that witnesses to unacknowledged past trauma.

Keywords

Memory mapping, quantum measurement, seen and unseen, transgenerational entanglement, trauma

Introduction

> When a pair of entangled particles is observed, the entanglement will be broken.
>
> (Mahood, 2018: 198)

Memory represents an entanglement with the past. When Syria's President Bashar al-Assad invokes the Armenian Genocide, or Hungary's Prime Minister Victor Orbán refers to the 'Ottoman invasion', they point to a traumatic past with political implications in the present. Arguments about

Corresponding author:
K. M. Fierke, University of St Andrews, School of International Relations, Arts Faculty Building, The Scores, St Andrews, Fife KY16 9AX, UK.
Email: kf30@st-andrews.ac.uk

collective trauma highlight the extent to which the narration of trauma expresses the concerns of successor generations, which may or may not be linked to an underlying traumatic experience of contemporary subjects (Alexander, 2012; Alexander et al., 2004; Sztompka, 2000). In this respect, memory is put in service of a present and future political project, which may be a source of community-building (Hutchison, 2016) and/or aggression (Fierke, 2004; Scheff and Rezinger, 1991). As Edkins (2002, 2003) notes, while trauma is 'unspeakable', given that it entails the rupture of everyday safety, it is often domesticated in political discourse.

The political uses of memory often hark back to events that long precede the living, which raises a question of why they would continue to exercise an affective pull on contemporary populations. Lerner (2019) argues that the affective resonance of the past is in part a function of the severity of the trauma or its recurring experience by communities over time. In this respect, the political discourse may express multiple traumas and layers of entanglement that continue to resonate with populations. For instance, the literature on historical trauma emphasizes the continuing impact of past traumas of a political nature on the present health of indigenous communities in particular, arising from interrelated genetic, social and environmental factors (Matthews and Phillips, 2010; Walters et al., 2011), as well as the continuing impact of structural violence on successor generations (Kirmayer et al., 2014: 311).

Starting with a claim that trauma represents an entanglement with the past, this article seeks to explore the quantum notion that to 'see' an entanglement is to break it, as well as the implications of this claim for the politics of security. While it is often assumed that the quantum debate in international relations remains at the level of theory (Lamb-Books, 2016), we explore the meaning of measurement, seeing and breaking an entanglement in the context of an 'experiment' regarding the ongoing impact of traumatic political memory on the present. How does quantum measurement differ from classical? What does it look like, and what are the implications for the analysis of security practice? What does it mean to say that when an entanglement is 'seen' it is broken? The analysis that follows arose from collaboration over the past four years between the two authors, one a scholar of international relations, the other a therapeutic practitioner with training in medical physics.[1] Our 'experiment' was motivated by an interest in exploring methods that approach problems of security from an angle different from that of conventional wisdom, thereby opening up new potentials, in the light of increasing questions about the efficacy of existing practices and a 'crisis' in the field (Nyman and Burke, 2016).[2]

In this article, our objective is to explore a conceptual problem rather than the experiment itself, or the results arising from it, not least owing to the difficulty of communicating an experiential method in language. In the first section, we explore a broad contrast between classical and quantum measurement, asking what this might mean at the macroscopic level. In the second section, we categorize Wendt's claim about language as a form of *expressive* measurement and explore the relationship to discourse analysis. The more spatial 'cut' into political discourse, while often drawing on memory, does not account for the affective resonance of transgenerational entanglements. In the third section, we explore the broad contours of our experiment and the role of a somewhat different form of non-linear expressive measurement. In the final section, we elaborate the relationship between *redemptive* measurement and breaking an entanglement, which involves a form of 'seeing' that witnesses to unacknowledged past trauma.

Classical and quantum

In classical physics, a particle can only be a particle and thus an independent entity, which gives rise to assumptions of materialism, locality and determinism. One of the central discoveries of quantum physics, as demonstrated in the famous two-slit experiment, is that a particle can become

a wave and a wave can become a particle in certain circumstances. Elementary particles are not objective material objects, with characteristics that can be determined, but rather phenomena that arise from an interaction of some kind, or indeed can be seen as the interaction itself. The latter contrasts with classical assumptions that an elementary particle is an independently existing entity. Instead, as the American quantum physicist Henry Stapp (1971: 1303) stated, a particle is 'in essence a set of relationships that reach toward other things'.

In classical physics, the separateness of objects with pre-existing properties and boundaries makes it possible to measure their interactions. There is an assumed intrinsic separation between the knower, the known and the apparatus of measurement itself. The scientist stands outside of the objects of observation, which are assumed to exist as discrete entities, with a fixed location in time and space. In simple terms, the measurement consists of quantifying the distance between objects. By contrast, in quantum physics the object of measurement is not fixed; the boundary that separates the object from the 'agencies of observation' will be heavily dependent on the physical arrangement of the apparatus, and thus indeterminate (Barad, 2007: 114). The apparatus is a crucial part of the measuring process. The choice of apparatus for each measurement creates the condition 'to give meaning to a particular set of variables at the exclusion of other essential variables' (Barad, 2007: 113–115). The apparatus and the measurement are entangled, and thus not entirely separable.

As Karen Barad notes (2007: 74), 'entanglements' are very specific configurations. However, it is difficult to build apparatuses for their study, because the apparatus changes with each intra-action, and because space, time and matter 'do not exist prior to the intra-actions that reconstitute entanglements'. The apparatus and the observed phenomenon change alongside one other. The measuring apparatus itself enacts a 'cut', which is an 'intra-action' from which separation and difference emerge (Barad, 2007: 140).[3] The intra-action between object and apparatus are a part of the phenomenon, which means that measurement practices also constitute the results and are thus indispensable to them. The analyst cannot be separated from the apparatus of measurement, and the measurement itself arises from an act of seeing.

The question is what form the apparatus would take in relation to human intra-actions. Barad conceptualizes the apparatus in broad material-discursive terms, which can take a variety of forms. Wendt (2015) specifies that language itself is an apparatus; language use is a form of measurement that impacts on what is observed. He states that 'in language what brings about a concept's collapse from potential meanings into an actual one is a speech act, which may be seen as a measurement that puts it into a context, with both other words and particular listeners' (Wendt, 2015: 217). The collapse starts with communicative intent (the decision to communicate one meaning rather than another), which depends also on the listener, whose understanding will depend on how what is said interacts with a memory of words and their association. Accordingly, 'memory structures relate to concepts in the same way that measurement devices in physics relate to particles', which suggests that quantum entanglement and interference are manifested in actual language use (Wendt, 2015: 217). Insofar as memories are stored not as isolated entities but as networks of related words, their entanglement is evident in how they are activated (Wendt, 2015: 219). The act of measurement begins with an intentional act of language use, by one who reaches out relationally to another. Memory is the repository of meanings from which the specific measurement arises, as wave functions collapse into language, materializing one potential rather than another.

In Wendt's argument, language use is both an expression of entanglement and the point of departure for the enactment of multiple potentials. What does this look like in practice? To take one example, as discussed elsewhere (Fierke, 2017), the concepts of migrant, refugee and terrorist as applied in the larger context of the European 'migration crisis' are, from this perspective, relational and defined in contrast to those who 'belong'. None of these categories map neatly onto a subject

with an intrinsic identity; rather, these are thin identity categories that are superimposed on the thicker sense of self that the incomer carries with them from a place of origin. In the confrontation between host society and incomers, the use of language is a measurement that places people along a status hierarchy that determines the extent of their 'humanness'. The language already *contains* a measurement of the identity of particular groups of people as human 'like us' or as less than human and a potential source of danger. This, then, also becomes a measure of what we should feel, whether compassion or fear, and how 'we' should act toward 'them' – that is, whether they should be welcomed or refused entry, held behind barbed wire or a wall, stripped of their possessions, tortured, or even killed. The example reinforces Wendt's claim that language use results in wave-function collapse around one potential rather than others, thereby instantiating one reality rather than other possible realities.

Expressive measurement

Language expresses a form of 'seeing' by the observer as wave functions collapse. The seeing is 'partial'. The thin concepts of 'migrant', 'refugee' or 'terrorist' define the boundaries within which the 'other' is seen, and any one 'cut' creates a particular separation between 'us' and 'them'. Another example highlights the role of memory, making it possible to explore the 'partiality' of the cut and measurement from a somewhat different angle. In a lengthy interview in 2014, President Bashar al-Assad made an unexpected reference to the massacres of 1.5 million Armenians and identified the perpetrator as Ottoman Turkey. During the interview, Assad compared the Armenian Genocide of 1915 to the brutal killings of civilians in Syria today:

> The degree of savagery and inhumanity that the terrorists have reached reminds us of what happened in the Middle Ages in Europe over 500 years ago. In more recent modern times, it reminds us of the massacres perpetrated by the Ottomans against the Armenians when they killed a million and a half Armenians and half a million Orthodox Syriacs in Syria and in Turkish territory. (Sassounian, 2014)

Assad's words were articulated in the context of a dispute with Turkey and were intended to lash back at the Turkish government's hostile actions against the Syrian regime.[4]

The example highlights several points. First, Assad's use of language is a *measurement* of the war in Syria that enacts a particular kind of separation between terrorists and states, which is magnified by reference to a particular memory of brutality. Memory, in this reading, is an observational instrument by which a particular 'cut' is made. Assad's reference to 'terrorists', associated with Turkey, places them outside of Syria, thereby reinforcing his legitimacy as the leader of Syria, as well as his actions in defence of Syria's security. A discourse of terrorists and legitimate leaders represents a measure different from that, for instance, of a Syrian 'civil war'. As suggested by Ricouer (1990: x), the identification of a resemblance between things that would at first glance seem to have nothing to do with each other 'grasps together' and integrates scattered events into a single whole. This involves a degree of forgetting, and thus elements that are not seen. The narrative excludes other possible alternatives and is itself selective. Second, the memory constitutes a future that is only 'there' as a project to be realized (Kratochwil, 2018: 420). Assad claims the memory as his own and frames it in a particular way that enables him to heighten his own use of violence, thereby drawing power from it. The implicit logic not only becomes a form of forgetting, in the light of its selectivity, but justifies his present and future project to eliminate the 'terrorists'. 'Seeing' happens from a particular position in time and space, which is partial. The measurement *expresses* a particular 'cut' that shapes a relational world in the present, one of Ottoman terrorists and legitimate state actors. An earlier perpetration fuels a perpetration in the present. Assad's use

of language is itself a form of *expressive* measurement by which he 'sees' the world. Third, the measurement of past, present and future obscures the more complex, open field within which the memory remains alive – in all those who were forcefully displaced, died or are otherwise unseen in their suffering, both past and present. As one Syrian blogger in Aleppo stated, during the siege in 2016, the world was not 'seeing' their suffering.

Discourse analysis

Assad's measurement could be analysed with methods of discourse analysis. One might, however, question the added value of associating discourse with abstract ideas about the activity of waves. The quantum argument is that physical systems do not have definite properties *until* they are measured through memory, and that it is at this point of observation that something comes to life. Assad's terrorists become agents of genocide as he invokes a memory that then has further physical or material consequences. The act of giving meaning is a collapse into the physical properties of language. While the quantum argument is interesting, discourse analysts have engaged in language analysis for decades without reference to wave function.

The quantum angle is, however, important for beginning to think differently about what it means to measure and how this relates to 'seeing'. Language-based methods have often been cast as 'fuzzy', woolly headed and therefore unscientific (Laffey and Weldes, 2004), and even users may be reluctant to associate discourse analysis with any kind of measurement. Measurement is associated with quantification, which rests on classical assumptions of atomism, as well as an understanding of language as a mirror that more or less accurately reflects truth in the world. Approaching language and measurement from a quantum angle turns this logic on its head. In Wendt's (2015: 217) argument, language use involves a speech act, which is a measurement that puts words in context, by which they are collapsed from a potential meaning into an actual one. Discourse analysis is the empirical study of relational worlds embedded in meaning structures that have been manifested in the words of political agents.

Discourse analysis is not just concerned with the mapping of relational worlds but, given its roots in Foucault, among others, has been particularly concerned with power relationships embedded in language. The analysis provides a means to 'see' the discourse not as a description of reality 'as it is', but as expressing a structure of power and exclusion. To take another example, which involves an even more problematic conflation, Hungary's Prime Minister Victor Orbán stated that incoming migrants represent an 'Ottoman invasion'.[5] The claim relates to two contrasting forms of 'seeing'. In the first, Orbán, like Assad, manifests a particular reality by invoking a specific memory. In this 'seeing', a population, composed primarily of people fleeing violence and persecution, becomes an invading army. The single claim is embedded in a larger relational world that is meaningful precisely because of the memory it brings to life. To 'see', in this use, is to go beyond a descriptive understanding of language to its embeddedness in relational structures of power.

A second form of 'seeing' arises from the *analysis* of the political statements of, for example, Orbán. From this position, we also begin to see not just the multiplicity of relational potentials but also the silences contained in discourse. The analysis might, for instance, juxtapose the 'invasion' with other measurements, for instance in Germany, that constructed a different world, characterized by the importance of compassion, relying perhaps on a memory of the plight of German refugees following World War II (see Feindt, 2017). A discourse analysis might also examine the power relationships inherent in either, arguing, for instance, that, even at its most humanitarian, the underlying logic is exclusionary and dehumanizing and thus silences the voices of the refugees themselves (Chouliaraki and Stolic, 2017; Musaro, 2017). As Chouliaraki and Stolic (2017: 1170) argue, the voicelessness of refugees is a form of 'epistemological violence' (Paik, 2016), in

which the marginalized become entangled with Western practices and discourses that reinforce their own exclusion. The analysis makes it possible to begin to 'see' the unseen, along with the suffering constituted by the discourses that surround refugees and migration, thereby breaking the hold of an entanglement.

Placing discourse analysis in a quantum framework reinforces several arguments that have been more or less successfully employed by critical scholars in the past.[6] While our analysis has focused on single texts of leaders, discourse analysis usually looks across the texts of multiple actors and thus tries to reconstruct a relational world. Discourse analysis does not by definition deal with memory, but it has been employed by analysts concerned with memory. A fourth generation of memory studies in the field of history, focused on 'entangled memory', represents a shift toward an emphasis on entanglement in discourse across time (see Feindt et al., 2014; Pestel et al., 2017).

The latter, nonetheless, remains limited by the availability of texts and, particularly when looking across time, the absence or destruction of documents or archives, not least relating to those who suffered and are unseen. For instance, a recent BBC documentary (Haymen, 2018) contrasted the myth of Scottish innocence in the slave trade, as well as Scotland's status as victim of England, with the many ways in which Glasgow, no less than Liverpool or Bristol, was closely bound up in and profited from the slave trade. This history, it was suggested, was written with the intention of not 'seeing' and expresses a national amnesia that was wilful and deliberate, written from the perspective of elite white men who controlled the archives, diaries and ledgers, which were often destroyed. National amnesia and forgetting worked at the level of a system that worked to erase, complemented by a public narrative in which all could participate in the obfuscation of 'reality'. In what follows, we suggest a method that is compatible with discourse analysis but goes further to explore the affective resonance of memory. In other words, in addition to understanding memory as an observational instrument by which a particular 'cut' is made, memory can be examined as itself an entangled phenomenon.

As already stated, public narratives are usually written from the perspective of present concerns and future projects (see also Kratochwil, 2018). This raises a question about the relationship between political discourse and entanglement with the past. How is it possible that battles that took place centuries ago have a continuing resonance in the present? Is this resonance on some level prior to discourse and entangled with an experiential past? In the next section, we highlight another form of the expressive measurement that is dependent on quantum effects, and thus on the relationship between collapsing wave functions and language.

Measuring transgenerational trauma

While potentially of tremendous scientific and practical significance, the experiential *and* experimental aspects of our project are difficult to communicate in words, not least owing to the nonlinear nature of the phenomenon and the divergence from conventional social science practice. The emphasis on the experiential as well as the experimental highlights the quantum assumption that the analyst or any participants cannot be separated from either the apparatus of measurement or the outcomes. 'Experience' in this case can be contrasted with both third-party 'objective' experience, most often associated with Newtonian science, and 'subjective' first-person experience, which Wendt (2015) discusses as consciousness of 'I'.[7] Instead, it can be thought of as a form of second-person experience that arises from an *interaction* between world and body, or with what Barad (2010: 260) refers to as memory that is 'written into the fabric of the world'. Our concern was less with the historical detail of what happened than with a diffracted relational pattern of affect that is entangled in memory, which we sought to map. The individuals involved engaged with the affect

surrounding a temporal phenomenon – that is, an experience that occurred in past time, rather than an entity, individual or otherwise.

Given space limitations and the primary intention to articulate a conceptual relationship between seeing and breaking an entanglement, we briefly present what the method is about and the relationship between the apparatus and forms of measurement involved. In doing so, we draw on insights and data from the experiment anecdotally to make conceptual points, and minimize the review or references to other literatures, except as necessary, given the numerous connections that could be made across fields in both the natural and the social sciences. While we recognize that the account may raise more questions than it answers, the unpacking of further concepts will have to wait for another time.

The method relies on several quantum assumptions. First, as already suggested, entangled phenomena are by definition *non-local*, and in this case express entanglements with transgenerational memory. The method takes a step beyond existing non-linear approaches to memory[8] to focus on the wave-function collapse itself and a quantum understanding of time. Barad (2014: 171) draws on the imagery of light behaving as a fluid, which, upon encountering an obstacle, breaks up and moves outward in different directions. Time itself is diffracted, she argues, insofar as it is 'broken apart in different directions, non-contemporaneous with itself. Each moment is an infinite multiplicity' (Barad, 2014: 169). Patterns of diffraction, as noted by Donna Haraway, do not mark where differences occur but rather where the *effects* of differences appear.[9] Our experiment shifts focus slightly to patterns of *affective* difference that emerge from the mapping of a relational whole, in which past and present are not fully separable.

Second, following on from the last point, entanglements relate to *emotions and affect.* As Sparrer (2007) notes, emotions do not 'belong' to us as stable attributes but can be both non-local and entangled with others, both present and past. The focus of our method is on traumatic entanglements with the past as they relate to political rather than individual memory.[10] It begins with an assumption that war, forced displacement and violence, suffered or perpetuated in one generation, cross over to other generations in such a way that a younger generation may bear the burdens of its parents' or grandparents' generations, thereby assuming the latter's unsanctioned behaviours and related guilt (Dietrich, 2013: 139). The main point is illustrated in a simple example at the individual level, where two men from different belief systems are embroiled in a deeply emotional fight, with each of them carrying the anger and experience of their father, grandfathers, great-grandfathers, etc. The entanglement with the past thus adds to the toxicity of current conflict.

A notion of transgenerational trauma points to a field of affective resonance that is beyond language. As Bessel van der Kolk (1998) noted:

> a century of study of traumatic memories shows that (i) semantic representations may coexist with sensory imprints; (ii) unlike trauma narratives, these sensory experiences often remain stable over time, unaltered by other life experiences; (iii) they may return, triggered by reminders, with a vividness as if the experience were happening all over again; and (iv) these flashbacks may occur in a mental state in which victims are unable to precisely articulate what they are feeling and thinking.

While Van der Kolk's focus is individual memory, we suggest that collective memories of trauma are entangled with sensory imprints – that is, the memory is itself an entangled phenomenon that can be triggered by political changes that bring past traumas to the surface.

Third, it is possible to map sensory imprints of transgenerational trauma – or what we refer to as fields of resonance. As we began our experiment, we relied on the basic principles and theory of systemic constellations therapy, which has established a central place for itself within German therapeutic culture, not least in efforts to address the traumatic after-effects of World War II (see

Bilger, 2016).[11] The basic idea of systems therapy more generally is that individual problems cannot be viewed in isolation from a larger relational system. While family systems therapy is often concerned with role-playing, systemic constellations go further, to the groundbreaking observation that it is possible, in certain circumstances, for substitutes or proxies – which we refer to as representatives – to experience the physical and affective dynamics of a system during a constellation exercise, thereby bringing insight into its deeper and often hidden affective dynamics (De Carvalho and Klussman, 2010). In other words, those who occupy positions within a relational system are able to represent the bodily sensations, feelings and impulses of someone whom they do not know or, in our experiment, that are associated with categories of memory that are larger than the individual. The phenomenon arises from an intra-action between diffraction patterns of affect, which express a collective experience of suffering in the past, and representatives within the experiment, who experience the affective resonance surrounding this past.

Similar to what many physicists have said about quantum physics more generally, Splinter and Wustehube state that the effectiveness of the systemic constellations approach 'can be regarded as empirically proven, but a broadly approved scientific explanation of *why* it works is missing' (Splinter and Wustehube, 2011: 118, emphasis in original). German psychologist and engineer Peter Schotter demonstrated in a scientific study involving 3000 individual experiences that the perceptions of proxies, who knew nothing of the parties they represented, were not random and were reproducible (De Carvalho and Klussman, 2010). One objective of our project was to determine whether, when moving from individual to political memory, patterns would emerge from the engagement of the representatives during the mapping process. The maps were set up *blind*, to minimize interpretation by the individuals involved and to establish that any patterns could be attributed to a field of resonance.[12]

Fourth, the individual and social or political dimensions of memory cannot be neatly separated, but the latter is the prior condition for the former. As Kratochwil (2018: 328) notes, individual memory is built up through participation in communication processes, which involve common reflections on who 'we' are, which is shaped by where we think we come from, none of which can be separated from identities and collective memories that make 'society' an ongoing and transgenerational concern among its members. While constellation work revolves around individuals, any one of whom will be entangled with diverse collective memories through their family lineage, the current project seeks to explore collective memory as prior to any one individual. Some work has been done to apply systemic constellations to political conflict, working directly with actors on the ground (see, for example, De Carvalho and Klussman, 2010; Mahr, 2003; Mayr, 2012; Splinter and Wustehube, 2011). While further development of this potential is very important, the present experiment began with an assumption that many current conflicts, such as that in Syria, are too hot or too dangerous to contemplate any direct engagement or, in the case of terrorist violence, also too difficult to address through direct involvement of the parties themselves. This gave rise to a conclusion that, as entangled memories arise from an experience of past generations, it is possible to represent categories of actors in a mapping exercise without going to the physical site of conflict, thereby minimizing risk factors for those involved while potentially bringing great benefits to those who are subjected daily to a diet of brutal violence.

As we worked further with the method, the frequent recurrence of memories of *forced displacement* led to a further distinction. The constellations are concerned with belonging and who belongs to a system, which in the case of political constellations is concerned with large groups (Dietrich, 2013: 134). The central importance of belonging, and of attachment to and having place within a group, highlights the extent to which forced displacement – as distinct from conflict, which tends to solidify group boundaries – represents the hard case. While forced displacement involves movements of large numbers of people, it represents a shattering of place and belonging within a society.

This suggested the usefulness of shifting away from a focus on individuals or conflict per se to memories of migration and forced displacement in the past, to ask how these contribute to the reification of contemporary divisions of belonging and non-belonging.

Discourse analysis provides a method for looking at the representation of phenomena in political discourse. What we refer to as *dynamic entangled memory mapping* (DEMM), by contrast, involves the representation and mapping of relational patterns of transgenerational traumatic memory, which, it is assumed, remain entangled with the past and fuel the affect surrounding contemporary migrations, among other things. This brings Wendt's general observations about language back to the context of an experiment in which, consistent with quantum mechanics, measurement brings about a wave-function collapse, which is a by-product of asking a particular question and preparing the experiment in such a way that it can be answered. The quantum effects that arise from the relational map, or more specifically the patterned expressions of affect that emerge out of the mapping process, point to a non-local field of resonance that is microscopic, while having macroscopic effects.[13]

The method does not measure a thing with intrinsic properties but relational positions within a system, the shape of which is heavily dependent on how the intentional question is asked. In this respect, the intentional question *is* the apparatus in our experiment; it animates and becomes a lens through which, for instance, to understand why the refugee/migrant is seen or not seen, and may be distorted by memories of past trauma. The discourse analyst examines *representation in political language*. By contrast, within our experiment, the *representatives*, who occupy positions within the memory map, *express the relational field* surrounding a traumatic past as they engage with one another, thereby *manifesting a physical presence* of the past *in language.* Specific maps explore forms of entanglement within a field or intersecting fields in more depth, including the hidden dimensions, which are less obvious in contemporary political articulations, thereby providing an affective measure of a relational whole, including the relationship between the seen (the political articulation) and the unseen (the victims, past or present).[14]

The intentional question is the apparatus for setting up an initial map in order to examine the relationship between the different elements of a system. Representatives who occupy positions within that system express, through words, bodily movements or gestures, the affect they experience while 'standing in' for any one position. These articulations express a form of wave-function collapse and a pattern of diffracted entanglement, by which the attributes of the system become visible or 'seen' as the vibrational frequencies surrounding a particular space, and the affect that arises from it, are transformed into language and thus became available for analysis.[15] If language use is a measure of wave-function collapse, the language arising from a relational system becomes a measure of a non-linear historical trauma field.

The field arises from a particular 'cut', shaped by the apparatus – that is, the intentional question – which constitutes a particular relational whole, including the hidden dimensions and the unseen. The three-dimensional memory maps can be contrasted with the one-dimensional field expressed in political discourse. For instance, our EU refugee/migration crisis pilot study began with an overarching question about what was standing in the way of a compassionate response to the refugees. The mapping method made it possible to explore the interplay of entangled memories, some of which were less visible, in constituting the dynamics of the relational field. For instance, as in the public discourse in Hungary, the memory of the Ottoman invasions had an active presence in the maps. One might have expected the Holocaust to be the more dominant influence, given that it is far more recent than the Ottoman invasions. Despite real-time images of refugees packed into trains or being thrown food like animals, which were reminiscent of the Holocaust, the Ottoman invasions had a more prominent place in the mapping. While the Holocaust did come into play, its role was recessive, and pulled back to a much earlier memory, specific to Hungary, namely, the 1848–49 Hungarian revolution against the Habsburgs and Russia, during which tens of thousands

of Hungarian civilians participated in antisemitic actions. This memory then became the focus of a separate map.

The reconstruction of multiple interfacing memories goes beyond Orbán's one-dimensional account of the Ottoman invasions to identify the recessive influence and continuing power of memories of perpetration in fuelling the emotional response to incoming migrants and refugees, even while the surface narrative is one of being a victim. Orbán's repeated comparison of the arrival of large numbers of refugees in Hungary to the Ottoman invasions is an acknowledgement of past suffering, but one that reinforces a division of pain that highlights Hungary's past as a victim. Memories of perpetration are more likely to occupy a recessive space and, precisely because they are hidden, to fuel present perpetration. In this respect, a complementary relationship between perpetrator and victim, in which the two are entangled across time, becomes evident. The spaces for migrants, the dead, the migrants' land of origin and those left behind were also encumbered with separate historical trauma fields. There were layers of memory upon memory, of which we only touched the surface. The added toxicity provided by these memories, and the distortion of the present they provided, will not have been helpful or positive for the healthy integration of incomers and would indeed severely hamper successful integration.

Language functions like the apparatus in a physics experiment, both for the political agents in real time and for the representatives in trauma time. The language of the intentional question, as formulated by the facilitator and case provider (e.g. What is standing in the way of a compassionate response to the refugees? What is standing in the way of delivery of aid to Aleppo? Or, Who can see slavery?), provides a cut that shapes the relational field of exploration, making it possible to discern relationships of belonging and not belonging. Rather than a linear statement of truth, the direct transcription of the words spoken by the representatives provides a non-linear record of a *conversation* between different parts of an observed system that expresses an entanglement with past experience. While the various conversations corresponded broadly to the historical record in question, they also revealed hidden dimensions that were contrary to the 'truth' as expressed in more accepted histories, which, as stated earlier, have often been written for purposes of 'not seeing'. It would also be feasible to employ a more conventional scientific apparatus to make quantitative measurements of the fields of resonance, measuring either changes in the brain frequencies of those who occupy positions within the maps or the changing frequency of the relational field itself. These forms of measurement are beyond the expertise of the authors but point to areas of potential collaboration with other disciplines.

Redemptive measurement

A further form of measurement provides a more human take on the quantum principle that an act of seeing breaks an entanglement, as well as the claim that measurement transforms the object of observation. What we refer to as a redemptive measurement involves beginning to see that which is hidden or unseen and to *give it a place of belonging* within the relational field. In this respect, there is a distinction between *expressive measurement* – that is, discourse analysis or the measurement of the non-linear conversation between representatives – and *redemptive measurement*, which, with the guidance of the facilitator, involves seeing, acknowledging and giving place to the unseen elements so that the traumatic entanglement is broken and a more positive relationality can begin to be restored. Redemptive measurement transforms a historical trauma field into a historical trauma *narrative*, in which the suffering is seen and the trauma loses some of its power.

Two distinct forms of language use constitute DEMM. The first is the spontaneous language expressed by the representatives within the map; the second is the more directed language narrative introduced by the facilitator. Unlike discourse analysis, which examines the partial view of political

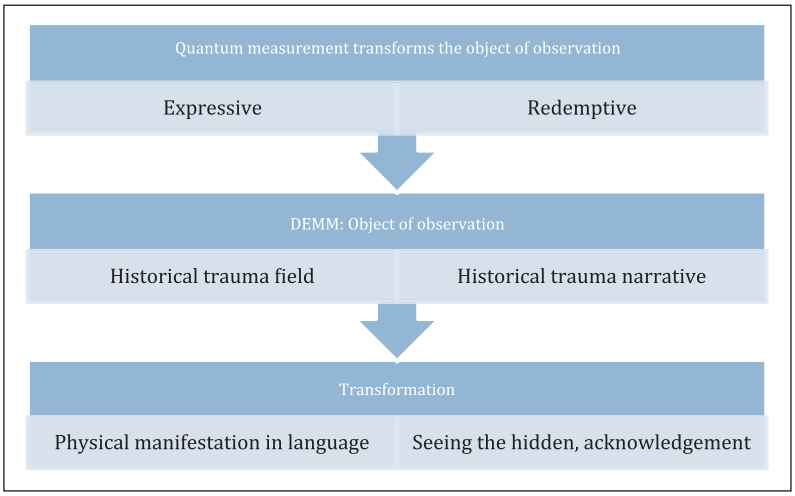


Figure 1. Quantum measurement and memory mapping. (Word Smart Art Graphic, with text provided by the authors.)

agents from the perspective of their present, the spontaneous language of the representatives expresses the relational whole, including the entanglements with historical trauma that shape the field of resonance. The more directed language of the facilitator, also informed by the intentional question, works with the representatives to change the narrative, thereby breaking the entanglement and provoking wave-function collapse around different potentials. The first expresses a field of habitual memory surrounding a past trauma; the second involves acknowledging and beginning to step outside the trauma, thereby paving the way for a different conversation.

A similar principle was expressed by, for instance, the South African Truth and Reconciliation Commission, where it was hoped that the trauma of apartheid would be lifted out of the individual experience of isolation and situated within a social narrative of the past, thereby restoring a sense of belonging within a relational whole, through the construction of a societal narrative.[16] While sharing this broad objective, the current project potentially addresses an ethical problem that has arisen in the context of truth and reconciliation commissions, namely, the estimated 50–60% of participants who were retraumatized as a result of participating in a public process (Hayner, 2001: 144). The other problem also encountered in the context of peace processes is that of getting those involved to meet face to face, given a highly toxic environment. The DEMM starts with a mapping of the trauma field and, in the process of measuring – that is, seeing and expressing a field in language – begins a dynamic process of transforming it into a historical trauma narrative that is redemptive, which, in theory, opens space for a different and less toxic conversation.

The potential impact of redemptive measurement is difficult to judge in the absence of a sustained experiment over several years that would also explore any relevant ethical questions in more depth. The issue here is whether in measuring and acknowledging the roots of traumatic memories in past suffering, an entanglement is broken and something changes in the world itself. The obvious answer, from the perspective of classical physics, is that it definitely would not have impact of this kind. However, on the basis of the quantum principle that measurement changes the object of observation, the DEMM could hypothetically have this impact – and indeed this is the purpose of the constellation method when applied in family and organizational therapy. The redemptive measurement of political memory opens a space for replacing the competition between conflicting memories with a broader conversation.

'Seeing'

But what, then, is meant by 'seeing' in this case? The contrast between the two forms of measurement highlights a number of issues. The first regards the *non-local dimensions* of both the transgenerational entanglements and the method. Expressive measurement transforms an unobservable field of resonance into language, where it can then be analysed in any number of ways. In the context of the experiment, the emergence of patterns from a *blind* process – that is, the expressive measurement – suggested that something powerful was going on, even while there is no clear explanation for why it works.[17]

Redemptive measurement is far more slippery insofar as it is difficult, if not impossible, to ever know for certain whether the measurement actually changes the object of observation and thus has broken the entanglement. For instance, one of the central themes that arose during the US Politics of Hate pilot study was the inability to 'see' slavery. After several days of working with a series of distinct maps, the representatives, who consistently, across separate maps, turned away from the occupant of the 'slavery' square, as though he or she wasn't there, began to engage with it in a way that had not been possible when we began. With guidance from the facilitator, the representatives began to acknowledge both the historical suffering and its continuing toxicity in that context. While any educated person knows the history of slavery in the USA, this knowledge is not the same as 'seeing'. To 'see', in this case, relates not only to acknowledgement of the suffering but also to entanglement within it. To 'see' is to witness. To witness is not merely a passive act of observation; it is rather an embodied act that makes the absence of memory present.

The quantum concept of complementarity might suggest that acknowledgement requires a recognition of the capacity for evil, as well as good, in any one self or community. From this perspective, the mutual implication of perpetrator and victim is more clearly evident where, as in the Hungarian case, memories of perpetration relating to antisemitism in Europe interfaced with memories of being victim during the Ottoman invasion. Or, in the case of the USA, its identity as a 'shining light on the hill' and a force for good in the world contains within it the barbarity of slavery (see Lepore, 2018). Free and equal US citizens, many of whom carried memories of persecution as immigrants from Europe, were the subjects of a constitution that emerged alongside laws regarding chattel slaves, who had been forcefully displaced from Africa. Once we begin to view the world from a different angle, recognizing ourselves as a part of life that is entangled across generations, as well as the planet, rather than standing outside and above it, the ethical bar for how we act toward others becomes much higher.

Balance in this conceptualization is not a mechanism, such as the balance of power, but an orientation to life, to self and the other in all its forms, of 'seeing' the humanity or more broadly 'seeing' life in the other, and of conversation with them (see Fierke and Jabri, 2019). While it may be tempting to regard this potential as utopian, particularly as regards the international, this misses the point. What is suggested is an ethical reorientation, and here a contrast is important. Many Western ethical systems rely on a metaphysics of atomistic rational individuals, for whom emotions are or should be absent and who are locally situated in time and space, in which the world is a mechanism and time is a quantitative measure, conceived in terms of clocks. By contrast, complementarity rests on quantum assumptions that the world is life, time is entangled, and affect is fundamental to life, including our humanity, and cannot be separated from reason. Insofar as the latter highlights the claim that harm done to others is ultimately harm to the self as well, it is consistent not only with Buddhist or African Ubuntu philosophy but also with a feminist ethic of care. But here we want to emphasize what this suggests about acknowledgement and the potential for redemption, and why both would be important. To redeem in this case is to acknowledge the reproduction of

harm within a particular relational system or structure, which is a first step toward re-establishing balance and rethinking of our security in relation to entangled others.

The second question regards *who precisely is better able to 'see'* as a result. The most straightforward and understandable answer would be that those who participate in the mapping process begin themselves, as a result of their representation of parts within a whole, to 'see' the previously unseen. They carry the experience as subjective witnesses during the mapping away from the exercise. While there was indeed evidence of this, as expressed by participants, even a year later,[18] the potential may extend further. For instance, in the weeks following the conclusion of the US Politics of Hate pilot study, the dramatization of far-right toxicity in Charlottesville, VA, made it impossible to ignore the continuing impact of a history of not 'seeing' slavery on contemporary politics in the USA. There was a corresponding emphasis in the media on the need for a conversation around the history of slavery. Was this increased ability to see slavery at all related to our experiment in a living room in Scotland? Any kind of causal claim about the relationship between redemptive measurement within the mapping exercise and changes in the world would be premature.[19] One objective of a longer, more sustained experiment would be to track the mapping in relation to multiple changing empirical contexts over time.

A third issue regards the potential for DEMM to have *a broader impact* on the world. If there were to be a larger impact, it would take the form of greater attention to the ongoing effects of structural violence and, in theory, a reduction in fear and the toxicity attached to the possibility of engaging in conversation about it. The greater ability to 'see' as a result of redemptive measurement may relate to those who participate in the mapping, thus making it a potential educational tool or a tool for engagement around policy – for example, examining the transgenerational entanglements between the descendants of European immigrants to the USA, including slave-owners, and those forcefully displaced from Africa.[20] Or, there may be more non-local and difficult-to-gauge changes, such as those suggested by the Charlottesville example. As physicist John Wheeler noted, in a quantum world we are 'participants in creating the universe', which requires that we take responsibility for it (Folger, 2002). As entangled participants, any transformation potentially impacts on us all. As suggested in relation to the contemporary 'crisis', refugees and migrants may themselves be the obvious victims, whether of conflict or of a lack of compassion by host societies, but the latter, and not least the United States or European countries, are also impacted by the failure to live up to their own core values, a failure that, we argue, is as entangled in memory as the former. Redemption points to our humanity and the ability to see the human in the other, not only in the present but in the past as well, and the other in the self. Redemption is less about changing the past (see Wendt, 2015) than about acknowledging it in such a way that we are changed in the now.

Conclusions

DEMM does not measure the distance between 'things' but rather the relationship between positions within a trauma field that is heavily dependent on the apparatus or how the intentional question is asked. The method makes it possible to 'cut' into the relational dynamics of specific historical dislocations to make visible the otherwise invisible affective resonance of transgenerational memory and its continuing impact on the present. There remains a question of whether this is primarily useful as a tool of analysis, for gaining insight into some of the hidden dimensions of memory on contemporary politics, or whether the non-local witness to and acknowledgement of historical patterns of displacement and perpetration changes something in the world itself. The mapping process makes it possible to measure a historical trauma field in expressive terms. Out of this process, the hidden dimensions of suffering relating to past trauma, which continue to impact

on the present, begin to surface. The corresponding concept of redemptive measurement suggests that in the process of 'seeing' and acknowledging this suffering, the entanglement is broken and something in the world itself changes, which is consistent with the quantum principle that observation changes the object of measurement. To 'see' is to break an entanglement. Beginning to see the other in the self, and the self in the other, including as this relates to victim–perpetrator dynamics across time, provides the basis for an ethical reorientation toward both the study and the practice of security.

Acknowledgements

This article was written by K. M. Fierke but is a product of collaboration with Nicola Mackay. The authors would like to thank the Centre for Peace and Conflict Studies at the University of St. Andrews and the Netherlands Institute for Advanced Studies for support for various aspects of this project, as well as Sofiia Bairamukova, Olga Burkhardt-Vetter, Roxane Farmanfarmaian, Caron Gentry, Frazer McDonald Hay, Naomi Head, Tony Lang, Mike Shanks, and Nadine Voelkner.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


K. M. Fierke https://orcid.org/0000-0001-6817-5171

Notes

1. We sought to apply a method that is widely used in the analysis of family and organizational systems (see, for example, Mackay, 2012; Roevens, 2008), particularly in Germany, to political phenomena. While Mackay has been a practitioner of the former for 17 years, Fierke's interest in the method was sparked by the quantum effects that arise from the dynamics of the systems analysis, which have no explanation in classical physics.
2. The word 'experiment' is used loosely to refer to the exploration of the usefulness of a particular method developed for one purpose to another. The process began out of curiosity, with no clear idea of where we were going, but it turned into three pilot studies relating to the delivery of humanitarian aid to Aleppo in 2016, the emergence of a US politics of hate following the 2016 elections and the EU refugee/migration 'crisis'.
3. Interaction assumes an exchange between separate parts, in which they remain unchanged. Intra-action, by contrast, begins with the whole and the constitution of separability as boundaries are drawn in an active process.
4. See Sassounian (2014).
5. Orbán here refers to Hungary's experience of conquest by the Ottoman Empire, going back to the 16th and 17th centuries. The decisive battle in the conquest of Hungary was the Battle of Mohacs in 1526, led by Sultan Suleyman the Magnificent, who defeated the medieval Kingdom of Hungary, which was far greater in size than the current country.
6. The quantum emphasis on measurement also highlights the distinction between a relational and an atomistic ontology, as expressed in the relationship between, for example, discourse and content analysis (see Herrera and Baumoeller, 2004). Further, it problematizes the relationship between observer and the apparatus of measurement. If the ability to 'see' is directly related to the apparatus of the observer – that is, their language use, which represents a 'cut' into a complex world – then there are obvious constraints on the degree to which claims of 'objectivity' can be made.
7. For a discussion of experience, including these two types, see Scott (1991).
8. In addition to the literature on entangled memory, non-linear approaches to memory would include, for instance, Rothberg (2009) or De Cesari and Rigney (2014).

9. As Haraway (1997: 273) further states, ‘Diffraction patterns record the history of interaction, interference, reinforcement, difference.’
10. This resonates with recent studies in epigenetics that suggest that traumatic entanglement can carry over from one generation to the next, altering genetic expression, while highlighting the role of environmental factors in triggering traumas of the past (see, for example, Daxinger and Whitelaw, 2016; Gapp et al., 2016).
11. The method, while influenced by forms of systems therapy, such as gestalt or Virginia Satir’s family sculpting, originated with Bert Hellinger (see, for example, Hellinger, 1999) and has become one of the most popular forms of therapy in Germany (see Bilger, 2016), although not without controversy, and has spread to some 25 countries.
12. In this respect, our project was more of a ‘pre-experiment’ to establish the validity of proceeding with a larger, more structured project.
13. This is contrary to frequent claims that quantum effects ‘wash out’ at the macroscopic level and are thus irrelevant for the social sciences; see, for example, Waldner (2017).
14. In the context of the US Politics of Hate pilot study, subject categories of a prior discourse analysis were used to set up the initial relational field of a specific map, which revolved around memories of the Civil War and slavery. This proved to be among the most powerful sessions, which suggests that use of the two methods in tandem may strengthen the results.
15. In the EU migration analysis, we constructed a literal transcription of the words spoken by the representatives as they moved around the mats engaging with one another. The completed transcription was then broken down into predicates – for example, subjects, verbs, adjectives, objects – and analysed in a manner similar to the way in which political discourse might be analysed.
16. The African concept of Ubuntu had an impact on the South African Truth and Reconciliation Commission, and Hellinger, who developed the constellation method, spent 16 years as a priest in South Africa observing indigenous healing practices; see, for example, Washington (2010).
17. While this may seem like what Einstein referred to as ‘spooky action at a distance’, other forms of ‘spooky action at distance’, from mobile phones to the internet to Skype, which at one time seemed a bit scary, are now a part of daily life. Like the memory mapping, these are all non-local phenomena but, unlike it, rely on technology.
18. For instance, following the presentation of a very moving paper that examined the testimony of a slave in the USA during the 19th century, Head (forthcoming) suggested a relationship between her participation in the US Politics of Hate pilot study and her later choice of this case, which was a divergence from her work on Israel–Palestine.
19. And would necessarily rest on a more non-linear understanding of causality; see, for example, Kurki (2008).
20. Mackay has been working with groups in Madison, WI, to this effect.

References

Alexander JC (2012) *Trauma: A Social Theory*. Cambridge: Polity.

Alexander JC, Eyerman R, Giesen B, Smelser NJ and Sztompka P (2004) *Cultural Trauma and Collective Identity*. Berkeley, CA: University of California Press.

Barad K (2007) *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.

Barad K (2010) Quantum entanglements and hauntological relations of inheritance: Dis/continuities, space-time enfoldings, and justice-to-come. *Derrida Today* 3(2): 240–268.

Barad K (2014) Diffracting diffraction: Cutting together apart. *Parallax* 20(3): 168–187.

Bilger B (2016) Where the Germans make peace with their dead. *The New Yorker*, 12 September.

Chouliaraki L and Stolic T (2017) Rethinking media responsibility in the refugee ‘crisis’: A visual typology of European news. *Media, Culture & Society* 39(8): 1162–1177.

Daxinger L and Whitelaw E (2016) Transgenerational epigenetic inheritance. *Genome Research* 20(12): 1623–1628.

De Carvalho M and Klussman J (2010) *Konfliktbearbeiten in Afghanistan. De Systemische Konfliktransformation in praktischen Einsatz bei einem Grossgruppen-konflikt* [*Conflict Processing in Afghanistan: Systemic Conflict Transformation in Practical Use in a Large Group Conflict*]. Berlin: Friedrich-Ebert Stiftung.

De Cesari and Rigney A (eds) (2014) *Transnational Memory: Circulation, Articulation, Scales*. Berlin: De Gruyter.

Dietrich W (2013) *Elicitive Conflict Transformation and the Transrational Shift in Peace Politics*. London: Palgrave.

Edkins J (2002) Forget trauma: Responses to September 11. *International Relations* 16(2): 243–256.

Edkins J (2003) *Trauma and the Memory of Politics*. Cambridge: Cambridge University Press.

Feindt G (2017) From 'flight and expulsion' to migration: Contextualizing German victims of forced migration. *European Review of History* 24(4): 552–577.

Feindt G, Krawatzek F, Mehler D, Pestel F and Trimcev R (2014) Entangled memory: Toward a third wave in memory studies. *History and Theory* 53(1): 24–44.

Fierke KM (2004) Whereof we can speak, thereof we must not be silent: Trauma, political solipsism and war. *Review of International Studies* 30(4): 471–491.

Fierke KM (2017) Consciousness at the interface: Wendt, Eastern wisdom and the ethics of intra-action. *Critical Review* 29(2): 141–169.

Fierke KM and Jabri V (2019) Global conversations: Relationality, embodiment and power in the move towards global IR. *Global Constitutionalism* 8(3): 506–535.

Folger T (2002) Does the universe exist if we're not looking? *Discover*, 1 June. Available at: https://www.discovermagazine.com/the-sciences/does-the-universe-exist-if-were-not-looking (accessed 4 January 2020).

Gapp K, Bohacek J, Grossman J, et al. (2016) Potential of environment enrichment to prevent transgenerational effects of paternal trauma. *Neuropsychopharmacology* 41: 2749–2758.

Haraway D (1997) *Modest_Witness@Second_Millennium.FemaleMan© Meets_OncoMouse™: Feminism and Technoscience*. New York: Routledge.

Haymen D (2018) *Slavery: Scotland's Hidden Shame*. BBC Two Scotland, 6 November.

Hayner PB (2001) *Unspeakable Truths: Confronting State Terror and Atrocity*. New York: Routledge.

Head N (forthcoming) Sentimental politics or responding to structural injustice: The ambivalence of emotions of political responsibility. *International Theory*.

Hellinger B (1999) *Acknowledging What Is*. Phoenix, AZ: Zeig, Tucker & Co.

Herrera YM and Baumoeller BF (eds) (2004) Symposium: Discourse and content analysis. *Qualitative Methods* 2(1): 15–39.

Hutchison E (2016) *Affective Communities in World Politics: Collective Emotions After Trauma*. Cambridge: Cambridge University Press.

Kirmayer LJ, Gone JP and Moses J (2014) Rethinking historical trauma. *Transcultural Psychiatry* 5(3): 299–319.

Kratochwil F (2018) *Praxis: On Acting and Knowing*. Cambridge: Cambridge University Press.

Kurki M (2008) *Causation in International Relations: Reclaiming Causal Analysis*. Cambridge: Cambridge University Press.

Laffey M and Weldes J (2004) Methodological reflections on discourse analysis. *Qualitative Methods* 2(1): 28–30.

Lamb-Books B (2016) Book review: *Quantum Mind and Social Science*. *Perspectives: A Newsletter of the ASA Theory Section*, 30 June. Available at: http://www.asatheory.org/current-newsletter-online/book-review-quantum-mind-and-social-science (accessed 24 December 2019).

Lepore J (2018) *These Truths: A History of the United States*. London: W. W. Norton.

Lerner A (2019) The uses and abuses of victimhood nationalism in international politics. *European Journal of International Relations*. Epub ahead of print 17 May 2019. DOI: 10.1177/1354066119850249.

Mackay N (2012) *Between the Lines*. Abingdon: John Hunt Publishing.

Mahood K (2018) *Entanglement*. London: HarperCollins.

Mahr A (2003) *Konfliktfelder-wissende Felder: Systemaufstellungen in der Friedens- und Versohnungsarbeit* [*Conflict Fields - Knowing Fields: Systemic Constellations in Peace and Reconciliation Work*]. Heidelberg: Carl Auer-Systeme-Verlag.

Matthews SG and Phillips DIW (2010) Minireview: Transgenerational inheritance of the stress response: A new frontier in stress research. *Endocrinology* 151(1): 7–13.

Mayr FP (2012) *Consciousing Relatedness: Systemic Conflict Transformation in Political Constellations*. Saarbrucken: Lambert Academic Publishing.

Musaro P (2017) Mare Nostrum: The visual politics of a military–humanitarian operation in the Mediterranean Sea. *Media, Culture & Society* 39(1): 11–28.

Nyman J and Burke A (2016) *Ethical Security Studies*. London: Routledge.

Paik AN (2016) *Rightlessness: Testimony and Redress in U.S. Prison Camps Since World War II*. Chapel Hill, NC: University of North Carolina Press.

Pestel F, Trimcev R, Feindt G and Krawatzek F (2017) Promise and challenge of European memory. *European Review of History* 24(4): 495–506.

Ricouer P (1990) *Time and Narrative. Vol. 1*. Chicago, IL: University of Chicago Press.

Roevens JLM (2008) *Systemic constellations work in organizations*. PhD dissertation, University of Tilburg, the Netherlands.

Rothberg M (2009) *Multidirectional Memory: Remembering the Holocaust in the Age of Decolonization*. Redwood City, CA: Stanford University Press.

Sassounian H (2014) Syrian president finally recognizes the Armenian genocide. *Asbarez*, 28 January. Available at: http://asbarez.com/118921/syrian-president-finally-recognizes-the-armenian-genocide/ (accessed 24 December 2019).

Scheff T and Rezinger S (1991) *Emotions and Violence: Shame and Rage in Destructive Conflicts*. Lanham, MD: Lexington Books.

Scott JW (1991) The evidence of experience. *Critical Inquiry* 17(4): 773–797.

Sparrer I (2007) *Miracle, Solution and System: Solution-Focused Systemic Structural Constellations for Therapy and Organisational Change*. Cheltenham: SolutionsBooks.

Splinter D and Wustehube L (2011) Discovering hidden dynamics: Applying systemic constellation work to ethnopolitical conflict. In: Korppen D, Ropers N and Giessmann HJ (eds) *The Non-Linearity of Peace Processes: Theory and Practice of Systemic Conflict Transformation*. Opladen: Verlag Barbara Budrich, 111–125.

Stapp HP (1971) S-Matrix interpretation of quantum theory. *Physical Review D* 3(6): 1303–1320.

Sztompka P (2000) Cultural trauma. *European Journal of Social Theory* 3(4): 449–466.

Van der Kolk BA (1998) Trauma and memory. *Psychiatry and Clinical Neurosciences* 52(S1): S52–S64.

Waldner D (2017) Schrodinger's cat and the dog that didn't bark: Why quantum mechanics is (probably) irrelevant to the social sciences. *Critical Review* 29(2): 199–233.

Walters KL, Beltran RE, Huh D and Evans-Campbell T (2011) Dis-placement and dis-ease: Land, place and health among American Indians and Alaska natives. In: Burton LM, Kemp SP, Leung MC, Matthews SA and Takeuchi DT (eds) *Communities, Neighborhood, and Health: Expanding the Boundaries of Place*. Philadelphia, PA: Springer Science+Business Media, 163–199.

Washington K (2010) Zulu traditional healing, Afrikan worldview and the practice of Ubuntu: Deep thought for Afrikan/black psychology. *The Journal of Pan African Studies* 3(8): 24–39.

Wendt A (2015) *Quantum Mind and Social Science: Unifying Physical and Social Ontology*. Cambridge: Cambridge University Press.

K. M. Fierke is a professor of international relations at the University of St. Andrews.

Nicola Mackay is a family constellation facilitator, teacher and author with 17 years of scientific and therapeutic experience.

Foundations of Science (2020) 25:247–257
https://doi.org/10.1007/s10699-019-09586-8

Entanglement of Art Coefficient, or Creativity

Kyoko Nakamura[1] · **Yukio Pegio Gunji**[2]

Published online: 23 February 2019

Abstract
While entanglement is a phenomenon discussed in quantum theory, it can also be found in art. We propose to connect entanglement to art's most fundamental question: what is creativity? For example, Marcel Duchamp found the essence of the creative act in the "art coefficient," the difference and/or gap between the artist's intention and realization which is created. This paper locates the common sense understanding of entanglement in an inseparable whole that ensures difference between the intention and realization. Seeing the artistic act as actively designing entanglement within artistic production, we present examples of this from the work of the Japanese-style painter Nakamura, and present a concrete vision for an answer regarding the question of the nature of creativity.

Keywords Art coefficient · Entanglement · Heterogeneity · Gap · Internal measurement

1 Introduction: The Art Coefficient and Creativity

What is creativity? Marcel Duchamp, in his lecture "Creative Act (Duchamp 1957)," expressed creativity as the art coefficient. The art coefficient is the difference between the artist's intention and realization which is created. For his 1917 *Fontaine* while Duchamp planned to express "Fontaine", a urinal was exhibited. A urinal is a ready-made product that cannot be anything but a urinal. Therefore, when a urinal called a "fountain" is presented at an art museum, art-experiencers are intensely jolted by the gap between the urinal and a fountain. In this gap enters the outside—for example, art-experiencers' interpretations. People have usually sought the significance of Duchamp's art coefficient in the happenstance encounter between artists and art-experiencers, and seen it as located in the unavoidability of the artwork being interpreted independently of the author's intention. However, the art coefficient can be seen as the difference between what one has planned and that which is realized in artistic production.

✉ Kyoko Nakamura
kyoko608@gmail.com

Yukio Pegio Gunji
pegioyukio@gmail.com

[1] Research Institute for Language and Cultures of Asia and Africa, Tokyo University of Foreign Studies, 3-11-1 Asahi-cho, Fuchu-shi, Tokyo 183-8534, Japan

[2] Department of Intermedia Art and Science, School of Fundamental Science and Technology, Waseda University, 3-4-1 Okubo, Shinjyuku-ku, Tokyo 169-8555, Japan

The plan can be defined not only to painting but also other arts and even to cognition and perception by referring to anticipation and/or post-diction. It is a strategy for understanding an outside that is formed actively yet in a way that is not completely controllable (Gunji 2018). The artist is, in other words, one who actively designs an art coefficient that cannot be deliberately managed (Nakamura and Gunji 2018).

An example of something with an art coefficient of zero is a plastic food model (Gunji 2018). Food plastic models exist so that customers can imagine in advance what the food offered is like, and they are expected to be as similar as possible to the food actually served. One can enjoy eating unexpected dishes in an *omakase* (i.e., chef's recommendation) course. In this case, a creative attitude that enjoys the kind of gap present in *Fontaine* is sought from both the cook and the customer. In creativity, gaps are welcomed and the greater the coefficient the greater the possibility of an impact and deep thought. With that said, not just anything with a large difference is good. That would be simple unrelatedness. The great gap between the intention and realization can entail the alternative of the two. For example, if you display a urinal as a "fountain" to some of your friends who are not familiar with arts, they are surprised at the gap between a urinal and a fountain. Since a urinal and a fountain cannot coexist, they can be alternative to each other. Sometimes it is regarded as a urinal, but sometimes it can be regarded as a fountain for your friends. If the gap for the experiencer is too great to allow for the concepts to be alternatives to each other, they will instead be assumed unrelated.

In the case of *Fontaine* exhibited by Duchamp, the urinal and the fountain are different in nature. While each concept is different, they are inseparable and exist together. With his work Duchamp set up a "coming and going" between the urinal and the fountain. By falling into this difference (gap), even art-experiencers—who are supposed to be unrelated to the production of the work—jump into and experience *Fontaine*, which they have been linked to without possessing any relationship to it. This kind of creation is the art coefficient. The core of this idea could appear in many other contexts. Since Austin (1962) proposes the idea of speech act, it is argued that the outside of a language is rushed into the language in the form of speech act. In our sense, various contexts and unpredictable meanings outside of the ordinal use of a word could rush into the gap between the speaker's intention and the listener's interpretation. Especially, the concepts of illocution and perlocution in linguistics is strongly relevant for our idea [e.g., (Cohen 1973)]. Many cognitive and/or linguistic illusion such as cognitive bias [e.g., Manktelow (2012)] and the liar paradox (Aerts et al 1999) could be interpreted in terms of the aspect of the art coefficient.

In quantum theory, entanglement refers to different pure states being tangled together (Gunji 2018). The art coefficient can be seen as entanglement in the artistic act. Therefore, artists show their skills in how they incorporate gaps into their work in an exquisite arrangement and how they introduce leaps to the outside into it. Below, we will present examples from three works of Nakamura, one of the authors of this paper, that give rise to artistic entanglement. Before presenting the examples, we explain the metaphorical notion of entanglement in the next section.

2 Metaphorical Notion of Entanglement

The notion of entanglement is used to describe states that are separated with each other but remain connected in a non-local way. Therefore, an entangled pair of two sets of states which is assumed to be expressed as a product of two sets that is strongly correlated.

We expand the notion of entanglement in a broader sense. In linguistic philosophy, Wittgenstein (1963) denied the meaning of a word and a sentence, and proposed the idea of performative sentence. Kripke (1980) took after Wittgensteins idea, and showed that if the referent of the name, Kurt Gödel, is defined by the person who proved the incompleteness theorem (IT), and if it is clarified that Gödel himself did not verify IT and plagiarized the idea, then one can be faced with the situation in which the person who proved IT did not prove IT. That implies co-existence of "prove IT" and "NOT(prove IT)" in the form of the liar paradox.

In Kripke's argument, "NOT(prove IT)" is just one of the examples outside of "prove IT (i.e., conventional referent of Kurt Gödel)". Various possible referents outside of the name, Kurt Gödel, can rush into the scene of performing the name, Kurt Gödel. In our context, various possible referents in the underlying context can rush into the gap between speaker's intention and listener's interpretation. It results in a paradoxical sentence in which NOT(prove IT)" is superposed on "prove IT". In other words, the sentence is entangled with underlying context which can contain various possible referents, and that entanglement can entail co-existence (i.e., superposition) of a specific conventional meaning and the alternative meaning of the sentence in an extreme case. However, entanglement of a sentence and context is unclear because context cannot explicitly appear.

In this sense perlocution and illocution play an essential role to understand the significance of entanglement in linguistic phenomena. When a sentence and its underlying context is replaced by a speech and an act, respectively, one can find illocution in the form of entanglement. The illocution can also entail a paradoxical speech such as, "Please speak in a small voice here" in a loud voice.

This idea is taken after by cognitive linguistics especially proposed by Lakoff (1987), in which the prototype of a given meaning is located at the center of the distribution of possible meanings, and the marginal area of the distribution is connected to the opposite to the prototype.

While co-existence of A and NOT(A) is one example of co-existence of different meanings, referents and contexts, co-existence of various things can be attracted by the gap between speaker's intention and listener's interpretation. We expand the idea of entanglement to contain such a broader sense of co-existence.

Especially, the mirror neurons could be strongly relevant for the notion of entanglement. It was previously considered that the mirror neurons of a monkey and a human brain can fire to the same act of both him/herself or of others (Gallese et al 1996; Rizzolatti et al 1996; Cochin et al 1998, 1999). Thus, it is considered that the sociality can result from the mirror neurons (Gallese 2003).

However, the notion of "the same" act can be misleading. How can one identify the same act for one's own act and other's act? The act cannot be separated from its goal, derived feeling, emotion, and so on. How can one determine the boundary of a specific act? Nobody can do that, and mirror neurons also cannot do so. In fact, it is recently reported that mirror neurons can fire not only a specific visual image of others which represents one's own act but also the image which can be relevant for the goal, feeling and/or emotions derived from the act (Bonini et al 2013; Urgesi et al 2010; Iacoboni 2009). The latter implies that representation of a specific act is entangled with its derived information processing such as inferring goal, feeling and emotion.

In our sense, information processing at a certain level cannot be separated from that at a higher level in the brain. The latter information processing is regarded as the information processing outside of the former one. Since information processing is interpreted as the interaction between a sender and a receiver of the information, we can consider that

the outside of a specific information can rush into the gap between the sender's intention and the receiver's interpretation. That is nothing but the notion of art coefficient and/or extended notion of entanglement.

As well as the origin of sociality resulting from the mirror neurons in the sense of entanglement (Sobhani et al 2012), the gap between the artist's intention and realization can entail the entanglement of the masterpiece and the outside of it. That is creativity hidden in the masterpiece featured with the gap between the intention and realization.

3 A Buddha Bug "*Buddhaptera*" Praising a Rapidly Ascending *Platyplotus* and Glistening

Platypuses live in Australia. They have a flat beak, like ducks (Grant 1989). They look like moles, and swim with the agility of fishes. Furthermore, they have a single foramen, like reptiles and birds. Excretion, procreation, and the laying of eggs all happen through one hole. While they lay eggs like birds, the hatched babies are nursed by their mother, like mammals. This mixed up living thing is a strange harmonious unification of various aspects. Nakamura took a strong interest in the platypus because it is itself a mediator of heterogeneous elements—in this gap. She was motivated to create by her desire to express this heterogeneity of a platypus.

One summer, she saw a lotus growing in a pot at a temple. She was driven by the intuitive desire to mix this lotus with up a platypus. In reality, this could never happen in its home of Australia. However, for Nakamura a lotus and platypus pointed to a heterogeneity that foretold entanglement. She composed the work so that it would beckon them. However, if one just reconstructs a platypus and lotus as a mixed thing without giving it much thought, there is even the possibility of debasing the heterogeneity of the platypus by turning it into a strange monster. How can one bring out the heterogeneity of the platypus in a way that can ensure various patches as a whole, the platypus?

When she was struggling with regard to this, she saw a jewel bug nymph taking shelter from rain near a window. On its back was an unestablished iridescent marble pattern. All of a sudden, in it she saw a shape resembling a human. For Nakamura, this was more than an illusion of the eyes. It brought about in her an intuitive feeling that the phenomenon of seeing resulting from the interaction of an object and an observer. This was the perspective of an internal measurement. When we perceive something, we tend to adopt a perspective that is based on the separation between that an observer and an object. External measurement is observation that rises above to look over the world as a whole. It is objective, scientific representation. On the other hand, in the case of internal measurement, the observer exists within the world and cannot observe without influencing that which is being observed (Matsuno 2016; Gunji et al 1997a, b; Gunji and Toyoda 1997). In other words, in internal measurement the exterior (that which cannot be controlled) is also internal to observation, rendering ineffective—although not dissolving—the difference between the exterior and interior and mixing up different things. Therefore, the co-existence of and the gap between an unrelated bug and human become possible. This shares commonalities with finding the Buddha—if one holds that this mixing up is entanglement, if with the attitude of an internal observer one holds that the consciousness that sees a human in a jewel bug sees the heterogeneity of the platypus. A "buddha bug" descended to the gap between the platypus and the lotus. This is a strange connection. In a way the "*Buddhaptera* (buddha bug)" worships and glistens, and *Platyplotus* (2015) was able to splendidly jump (Fig. 1).

Fig. 1 *Platyplotus—Sudden rise* 2015 Kyoko Nakamura Color on silk, 170 × 68 cm (Overall view: left; Enlarged part: right above and right center) and *Buddhaptera* 2016 Kyoko Nakamura Color on silk, 15 × 15 cm (right below)

4 An Entangled Meal

At one point, Nakamura saw the movie *Moby Dick* (1956), based on Herman Melville's novel. In the painting, a whale arises amidst the flashing of lightning in chaos. This is the signature of the whale depicted by Queequeg. It is the large, white whale named "Moby Dick" that was never to be captured. Beckoned by Captain Ahab, the crew goes towards the whale, who is reaching his end. At first Nakamura wanted to paint such a whale. However, after reading *Le nouveau monde amoureux* (Fourier, translated in Japanese, 2013), the theme of whale suddenly escaped her, and was driven by the impulse to depict a meal in the harmonious world that Fourier claimed. What is called, combined harmonization could consist of individuals' various taste and happiness. Fourier's dream on new society is the attitude to enable this harmonization, which can be implemented by the internal measurement consistent with Fourier's terminology, "infinitely small (infinitésimal)". On one hand, the internal measurement perpetually mixes the measured state with external perturbation. On the other hand, the infinitely small enables harmonization of intrinsically different things. Both of them can enable happiness

Fig. 2 *Sawachi de Moby Dick* 2016 Kyoko Nakamura Handscroll, color on silk, 34 × 1423 cm

resulting from entanglement (Nakamura 2018). Although we do not refer to Fourier further in this paper, there is a more important discussion about the idea between Fourier and Nakamura's works. We would like to discuss it at another opportunity. Nakamura found *sawachi dishes*, a kind of local cuisine from Kochi (the name of a place) Prefecture in Japan made at home for guests on celebratory occasions such as weddings and birthdays (Matsuzaki 1986).

Eaten in large groups, the food is placed on a large plate with a diameter of around twelve inches or more. This plate, which brings together hors d'oeuvre, a main fish dish, side dishes, and dessert, is a highlight of *sawachi dishes*. This style of dining originates in *naorai* (a ceremony held at the end of a Shinto rite in which participants would share

food that had been offered to the god). This tradition has been passed down across Japan as a traditional feast in which the god and people eat together, strengthening their connections. However, due to its popularization, its original premise was lost, leading to difference between what is planned and what is realized. There is no order of dishes or hierarchy between foods as there is in course meals. Guests eat what they like when they want to. While going ahead with their own meal on their own terms, they deepen their relationships with others. Also, in *sawachi dishes* people—the old, the young, men, women—are equal around the plate. In Kochi, holding such feasts is referred to as *okyaku suru*. Both people from the family who invited others as well as those who were invited sit around the plate. Therefore, the relationship of "entertaining" and "being entertained," as well as the distinctions between leading and supporting roles, gradually becomes ambiguous. The meaning of the gathering in the end does as well, and it turns into just a drinking party. *Sawachi dishes* include lots of food with an emphasis on appearance: a plover-like arrangement of sea bream that appears to have gathered together in the ocean, and even bonsai tree-like plates. When the drinking party goes in full swing, people drink together while entertaining each other using chopsticks and cups, and guests become constitutive elements of the plate and fragmented, decorating the feast. Everything that comprised the meal becomes completely fragmented, and a style aspiring towards an infinitesimally small heterogeneity becomes a medium. Even the bites of mackerel sushi deliciously brought to one's mouth take on a divine nature. In note at the end of *Sawachi de Moby Dick* (2016) (Fig. 2), Gunji says, "A god's messenger-like frog and sparrow, who play the role of mediator, continually adjust a breaking set of things. While making the top of a plate be a background with the pieced together remaining offerings, each individual's meal is continually combined and opened up as a site of summoning (Gunji 2017)."

The head of a mackerel that remains on the plate during a party was like a whale raising its head from the surface of the ocean. At this time, for Nakamura, the whale descended to the gap of *sawachi dishes*. However, on the plate was not a great, large Moby Dick, but a life-sized one. This the entanglement of humans (food) and whales reflected by a small yet divine plate.

5 *Someone Digging Landscape—Known Againness*

In the early afternoon when it's still a bit chilly, one goes to the ocean shore at low tide. The bubbly sand feels nice and sticky on one's bare foot. Using the gentle slopes of the rolling beach as a guide, one dips one's legs in the sand, feeling the cold of the ground and something hard. It is a clam. Using one's big toe, one carefully gets it out of the sand, and takes into one's a hand a skyscraper (the pattern on the shellfish). While searching for various "landscapes"—peonies, running crows, snowstorms, piled umbrellas—one looks into the distance and finds the pattern in the shells of clam. Unknown views are more nostalgic than reality.

For Nakamura shellfish gathering itself is synonymous with nostalgic things. Just by hearing the phrase "shellfish gathering," she experiences a time that is more nostalgic than reality, like gazing out into an unknown distant scene. In other words, shellfish gathering is time. Is not the sense of time also manifested by entanglement? There is a tendency to view time as always being fixed, going in one direction: from the past to the future. However, in reality this is not the case. Gunji proposed an inference system equipped with both

Fig. 3 *Someone digging landscape—Known againness* 2017 Kyoko Nakamura Hanging scroll, color on silk, 109 × 88.5 cm

Bayesian and inverse Bayesian inference (Gunji et al 2016, 2017), showing the co-existence of "past experience/ 'I' of the past" and "present perception/ 'I' of the present (Gunji 2018)." Gunji sees this as generalized déjà vu. The everyday is filled with déjà vu. Time, which clearly extends the closer the memory is to one's earliest memory from childhood and is supposed to be fixed, is felt to be shorter and shorter as one goes back to the present (Nabokov 2017). Amidst coming and going between Bayesian inference and inverse Bayesian inference, we invite the exterior that is time into the interior. With them co-existing, we are always forming time. Is pure time not entanglement that always descends into the gap from the exterior?

In Japan there is the superstition that mirages arise from the energy released by shellfish. The distant scenery seen by Nakamura while searching for various landscapes was the clam that creates a tower on its body and turns into a landscape. This landscape that is more nostalgic than reality is time itself in which things of different natures things co-exist. Shellfish, which go back and forth between manifest imagination and latent reality, were the ones searching for landscapes (Fig. 3).

6 Conclusion: The Universality of Heterogeneous Things

The entanglement in Nakamura's artistic production discussed in this paper is very arbitrary. Throwing a platypus and lotus together, seeing a human in a jewel bug, a *sawachi dishes* feast facing Moby Dick, shellfish becoming a landscape—these are all only

meaningful for Nakamura. What is arbitrariness = heterogeneity for an individual only has meaning for that person. In principle, the meaning of arbitrariness hinders external evaluation (Gunji 2018). There is no universal value in arbitrariness, with respect to external measurement. The emotions of enjoying nostalgia and deliciousness are in the first place arbitrary and therefore completely different in nature from each other. One cannot share or exchange the same deliciousness. While different things cannot co-exist, exchange and fused, they can be harmonized because they can confront with each other. Confronting with each other is one of implementations of harmonization with respect to difference.

There is a major difference in "universality" depending on whether one sees homogeneity with invariance or sees heterogeneity with originality (Nakamura 2016). Because these emotions are different—unrelated and purely meaningless from the perspective of the world and only meaningful to this "I"—each person satisfies a universality that they can be confident is happiness. In the first place, was entanglement not the tangling together of pure states? This is the universality of things that are different in nature.

Heterogeneity in artist's creative acts, is a sensual metaphor. Even Duchamp did not yield a concrete image of artificial coefficient. We here develop concrete image of artificial coefficient, by introducing entanglement compared to the entanglement in quantum mechanics. Speaking in our terminology, entanglement may not be consistent with the entanglement in quantum mechanics, since we focus on macroscopic creative acts. For instance, it is not easy to estimate the virtual dimension in biological time in terms of quantum mechanics. Although our discourse could contain some issues on biological and/or psychological problem, we consider that the essence of entanglement in quantum mechanics is based on the mixture of different pure states or harmonization of different things, which can yield the essential mechanism of creative acts.

In starting from Duchamp's art coefficient, we propose the connection between entanglement and creativity resulting from the art coefficient. While entanglement in quantum mechanics represents correlated states which can be assumed to be independently separated with each other, we expand the notion of entanglement of which various contexts outside of a particular context can be attracted by the gap between the intention and realization in the particular context. Such an extended entanglement can be found in various fields, not only in cognitive linguistics but in brain science referring to mirror neurons. While we focus on the emergence of a masterpiece of art resulting from entanglement, the origin of sociality and/or language could be explained by the notion of entanglement. Various other notions that are part of the quantum epistemology such as contextuality, superposition, and emergence, might be useful to improve the notion of art coefficient and thus creativity (see also Gabora 2002).

Acknowledgements We acknowledge JSPS for the financial support of our work, Project Numbers 17K18465 and 18K18478.

References

Aerts, D., Broekaert, J., & Smets, S. (1999). A quantum structure description of the liar paradox. *International Journal of Theoretical Physics*, *38*(12), 3231–3239.

Austin, J. L. (1962). *How to do things with words*. Cambridge University Press.

Bonini, L., Ferrari, P. F., & Fogassi, L. (2013). Neurophysiological bases underlying the organization of intentional actions and the understanding of others' intention. *Consciousness and Cognition*, *22*, 1095–1104.

Cochin, S., Barthelemy, C., Lejeune, B., Roux, S., & Martineau, J. (1998). Perception of motion and qEEG activity in human adults. *Electroencephalography and Clinical Neurophysiology*, *107*, 287–295.

Cochin, S., Barthelemy, C., Roux, S., & Martineau, J. (1999). Observation and execution of movement: Similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience*, *11*, 1839–1842.

Cohen, T. (1973). Illocutions and perlocutions. *Foundations of Language*, *9*, 492–503.

Duchamp, M. (1957). *Creative act*. https://www.brainpickings.org/2012/08/23/the-creative-act-marcel-duchamp-1957.

Fourier, C. (2013). *Le nouveau monde amoureux*. Sakuhin-Sha, Pub. Co., Tokyo, Japan (Fukushima, T. translation to Japanese).

Gabora, L. (2002). Cognitive mechanisms underlying the creative process. In: *Proceedings of the 4th conference on creativity and cognition* (pp 126–133). ACM

Gallese, V. (2003). The roots of empathy: The shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology*, *36*, 171–180.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 593–609.

Grant, T. (1989). *The platypus: A unique Mammal*. Kensington: New South Wales University Press.

Gunji, Y. P. (2017). *Sawachi-dishes*. Note (Addendum) at the end of *Sawachi de Moby Dick* (in Japanese).

Gunji, Y. P. (2018). *Life do not move against artificial Intelligence*. Tokyo: Seido-Sha, Pub. Co. (in Japanese).

Gunji, Y. P., Ito, K., & Kusunoki, Y. (1997b). Formal model of internal measurement: Alternate changing between recursive definition and domain equation. *Physica D: Nonlinear Phenomena*, *110*, 289–312.

Gunji, Y. P., Rssler, E. O., & Matsuno, K. (1997a). *Internal measurement: Science of complex systems and modern thought*. Tokyo: Seido-Sha, Pub. Co. (in Japanese).

Gunji, Y. P., Shinohara, S., Haruna, T., & Basios, V. (2017). Inverse bayesian inference as a key of consciousness featuring a macroscopic quantum logic structure. *Biosystems*, *152*, 44–63.

Gunji, Y. P., Sonoda, K., & Basios, V. (2016). Quantum cognition based on an ambiguous representation derived from a rough set approximation. *Biosystems*, *141*, 55–66.

Gunji, Y. P., & Toyoda, S. (1997). Dynamically changing interface as a model of measurement in complex systems. *Physica D: Nonlinear Phenomena*, *101*, 27–54.

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, *60*, 653–670.

Kripke, S. (1980). *Naming and necessity*. Harvard University Press.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Manktelow, K. (2012). *Thinking and reasoning: An introduction to the psychology of reason, judgement and decision making*. New York: Psychology Press.

Matsuno, K. (2016). *Internal measurement coming*. Tokyo: Kodan-sha, Pub. Co. (in Japanese).

Matsuzaki, J. (1986). *Eating habits complete works 39, Meals in Kochi*. Minato: Rural Culture Association Japan (in Japanese).

Nabokov, V. (2017). Ada or Ador. (Wakahsima, T. translation to Japanese) hayakawa-shobo, Pub. Co., Tokyo, Japan (in Japanese).

Nakamura, K. (2016). How humans produce nature: The heterogeneous and the universal. In N. Kasuga (Ed.), *Bridging science with culture: Analogical thinking*. Tokyo: University of Tokyo Press. (in Japanese).

Nakamura, K. (2018). *L'archibras se relève*. Colloque « Fourier ! Fourier ! Deux journées avec Charles Fourier », Tokyo: Hitotsubashi University (in Japanese and French translation).

Nakamura, K., & Gunji, Y. P. (2018). *TANKURI: Shooting creativity*. Tokyo: Suisei-sha, Pub. Co. (in Japanese).

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*, 131–141.

Sobhani, M., Fox, G. R., Kaplan, J., & Aziz-Zadeh, L. (2012). Interpersonal liking modulates motor-related neural regions. *PloS ONE*, *7*, e46809.

Urgesi, C., Maieron, M., Avenanti, A., Tidoni, E., Fabbro, F., & Aglioti, S. M. (2010). Simulating the future of actions in the human corticospinal system. *Cerebral Cortex*, *20*, 2511–2521.
Wittgenstein, L. (1963). *Philosophical investigations* (translated by G.E.M. Anscombe) Basil Blackwell, Oxford.

Kyoko Nakamura is a "Japanese-style painting" artist. She holds a PhD in Fine Art from Tokyo University of the Arts, Tokyo, Japan. Her research background includes fundamental aspect of the creativity found in life, Human and Nature. She and Gunji published a book of her own paintings and discussed on creativity, titled "TANKURI" (published in 2018, Suisei-sha, Pub. Co., in Japanese). This book will be published also in English in 2019.

Yukio-Pegio Gunji is a professor of Department of Intermedia, Art and Science, School of Fundamental Science and Engineering, Waseda University, Japan. He graduated and got Doctor of Science at Tohoku university, Japan. He firstly studied geology and paleontology, and then shifted to theoretical biology and ethology. He proposed the idea of internal measurement by which an endo-observer can anticipate the outside of his/her own perspective. He published 10 books in Japanese and over 200 refereed journal papers.

APL Photonics

Entanglement distillation by Hong-Ou-Mandel interference with orbital angular momentum states

Cite as: APL Photonics 4, 016103 (2019); https://doi.org/10.1063/1.5079970
Submitted: 05 November 2018 • Accepted: 04 January 2019 • Published Online: 25 January 2019

B. Ndagano and A. Forbes

APL Photonics ARTICLE scitation.org/journal/app

Entanglement distillation by Hong-Ou-Mandel interference with orbital angular momentum states

Cite as: APL Photon. 4, 016103 (2019); doi: 10.1063/1.5079970
Submitted: 5 November 2018 • Accepted: 4 January 2019 •
Published Online: 25 January 2019

B. Ndagano[a] and A. Forbes

AFFILIATIONS

School of Physics, University of the Witwatersrand, Private Bag 3, Wits 2050, South Africa

[a]Electronic mail: nibienvenu@gmail.com

ABSTRACT

Entanglement is an invaluable resource to various quantum communication, metrology, and computing processes. In particular, spatial entanglement has become topical, owing to its wider Hilbert space that allows photons to carry more information. However, spatial entanglement is susceptible to decay in the presence of external perturbations such as atmospheric turbulence. Here we show theoretically and experimentally that in a weak turbulence regime, maximally entangled states can be distilled through quantum interference. We generated entangled photons by spontaneous parametric down-conversion, with one photon in the entangled pairs being sent through a turbulent channel. We recombined the paths of the two photons at a beam-splitter in a Hong-Ou-Mandel interference setup and measured in coincidence, using spatial filters, the spatial correlations between photons in the output ports of the beam-splitter. We performed a state tomography and show that, from an ensemble of pure states with very low levels of entanglement, we distil entangled states with fidelities $F \geq 0.90$ with respect to the singlet Bell state.

I. INTRODUCTION

Generating, manipulating, and sharing quantum states with maximal levels of entanglement are crucial steps to realising quantum processes such as quantum key distribution, teleportation, metrology, and computation.[1–11] In this quest, realising entanglement in different degrees of freedom has opened avenues beyond the two-level quantum bit (qubit). Spatial modes, particularly those carrying orbital angular momentum (OAM), allow one to exploit the spatial properties of photons to realise high-dimensional entanglement.[12–15] However, correlations between entangled spatial modes are adversely affected by external factors such as atmospheric turbulence.[16–22] These perturbations reduce the degree of entanglement and, consequently, the fidelity of the quantum process implemented. This constitutes a major hindrance in processes where maximally entangled states are required. Methods to mitigate the effects of turbulence on spatial modes have been proposed, with some notable ones including increasing the separation in mode space to reduce cross talk[23] and performing a coordinate transformation on one of the entangled photons to cancel out antisymmetric contributions of the turbulence.[24]

Through entanglement concentration or distillation, one can sift, from an ensemble, a fraction of states with a higher degree of entanglement.[25–27] In the special case of pure states, entanglement distillation can be realised in two ways. The first is through Procrustean filtering, where local operations on the entangled pair are performed in order to control the probability amplitudes of the post-selected states. Naturally, this requires prior knowledge of the quantum states before filtering. First demonstrated with polarisation states,[28] this technique has been extended to spatial modes in higher dimensions.[15] The second method is the Schmidt projection and, unlike the first, can be applied to an unknown quantum state. Although efficient for large ensembles of entangled

pairs, this scheme is impractical as it requires collective measurements to be performed on the ensemble.[25] Alternative schemes to realise the Schmidt projection have been proposed to circumvent the requirement for collective measurements, many of which involve ancillary photons (pairs).[29–33]

In the present work, we address the issue of entanglement distillation on an ensemble of OAM entangled photons generated by spontaneous parametric down-conversion (SPDC) and perturbed by a one-sided weak turbulent channel. One of the photons in the entangled pair propagates through the turbulence, leading to a unitary scattering in a higher dimensional OAM space. As a result of post-selection on a particular OAM subspace, one measures a decay of the degree of entanglement despite the unitary nature of the channel operator. To distil entanglement, we interfere the non-maximally entangled photons in a Hong-Ou-Mandel (HOM) configuration.[34,35] It has been shown that the HOM interference can be exploited to implement a filter for Bell states according to their intrinsic symmetry.[36–38] By performing a state tomography of the quantum states after implementing the HOM filter, we obtained distilled entangled singlet states with fidelity $F \geq 0.90$, up from as low as $F = 0.03$.

II. THEORY

Consider the OAM entangled two-photon state $|\Psi_\ell^-\rangle$ expressed as follows:

$$|\Psi_\ell^\pm\rangle = \frac{1}{\sqrt{2}}(|\ell\rangle_A|-\ell\rangle_B \pm |-\ell\rangle_A|\ell\rangle_B), \tag{1}$$

where the subscripts A and B label each of the photons in the entangled pair carrying $\ell\hbar$ quanta of OAM. Photon A is allowed to propagate through a turbulent channel that is unitary, causing a scattering of OAM states,

$$|\ell\rangle \xrightarrow{\text{turbulence}} \sum_{\ell'} c_{\ell-\ell'}|\ell'\rangle, \tag{2}$$

where $\sum_{\ell'}|c_{\ell-\ell'}|^2 = 1$. While the scattering certainly leads to the formation of higher-dimensional correlations, we will restrict the problem to a qubit subspace by projecting on the initial OAM state space; that is, we will consider the terms with $\ell' = \pm\ell$ in Eq. (2). The perturbed qubit state after turbulence then becomes

$$|\Psi_\ell\rangle = \frac{1}{\sqrt{2}}\Big(c_0|\ell\rangle_A|-\ell\rangle_B - c_0|-\ell\rangle_A|\ell\rangle_B + c_{2\ell}|-\ell\rangle_A|-\ell\rangle_B - c_{-2\ell}|\ell\rangle_A|\ell\rangle_B\Big). \tag{3}$$

It is useful at this point to express $|\Psi_\ell\rangle$ in the Bell basis. This can be done by realising that the last two terms in Eq. (3) correspond to a state on a two-photon OAM Bloch sphere,[39] where the poles are the Bell states $|\Phi_\ell^+\rangle$ and $|\Phi_\ell^-\rangle$, expressed as follows:

$$|\Phi_\ell^\pm\rangle = \frac{1}{\sqrt{2}}(|-\ell\rangle_A|-\ell\rangle_B \pm |\ell\rangle_A|\ell\rangle_B). \tag{4}$$

One then rewrites Eq. (3), the Bell basis, as follows:

$$|\Psi_\ell\rangle = c_0|\Psi_\ell^-\rangle + c_{2\ell}^+|\Phi_\ell^+\rangle + c_{2\ell}^-|\Phi_\ell^-\rangle, \tag{5}$$

where $c_{2\ell}^\pm = \langle\Phi_\ell^\pm|\Psi_\ell\rangle$.

After the turbulent channel, the photons enter a Hong-Ou-Mandel filter, as shown in Fig. 1(a). The aim of this filter is to distil the singlet state $|\Psi_\ell^-\rangle$ from the perturbed state $|\Psi_\ell\rangle$. The layout of the filter is shown as the inset in Fig. 1(a) and consists of two mirrors and a 50:50 beam-splitter (BS).

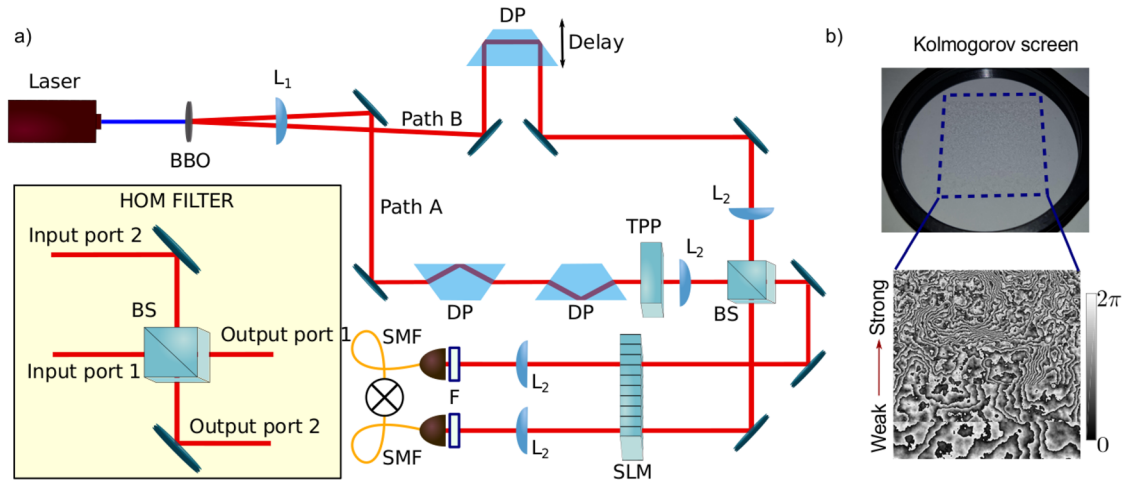


FIG. 1. Experimental setup for entanglement distillation. (a) By pumping a 3-mm thick, non-linear type-I BBO crystal with a 355 nm laser, we produced, through SPDC, pairs of entangled photons with wavelength 710 nm. The photons are sent down two paths, one containing a delay line and another with two Dove prisms (DP) to manipulate the SPDC state symmetry. A turbulence phase plate (TPP) in path A perturbs the state before the photon paths are recombined at a 50:50 beam-splitter (BS). Lenses L_1 and L_2 with focal lengths $f_1 = 100$ mm and $f_2 = 750$ mm, respectively, relay the plane of the BBO crystal onto the spatial light modulator (SLM). The two photons are probed using digital holograms encoded on the SLM, passed through 10 nm bandpass filters (F), and then coupled to single mode fibres (SMF). (b) shows the range of phase fluctuations across the turbulence phase plate.

A general entangled state $|\Psi\rangle = (|\ell_1\rangle_A|-\ell_2\rangle_B \pm |-\ell_1\rangle_A|\ell_2\rangle_B)/\sqrt{2}$ is transformed by the HOM filter as follows:

$$|\Psi\rangle \xrightarrow{\text{HOM filter}} \frac{1}{2\sqrt{2}}\Big(i|\ell_1\rangle^1|-\ell_2\rangle^1 + i|\ell_1\rangle^2|-\ell_2\rangle^2 + |\ell_1\rangle^1|-\ell_2\rangle^2 - |\ell_1\rangle^2|-\ell_2\rangle^1\Big) \pm \frac{1}{2\sqrt{2}}\Big(i|-\ell_1\rangle^1|\ell_2\rangle^1 + i|-\ell_1\rangle^2|\ell_2\rangle^2 - |-\ell_1\rangle^1|\ell_2\rangle^2 + |-\ell_1\rangle^2|\ell_2\rangle^1\Big), \tag{6}$$

where the superscripts 1 and 2 label the output ports of the beam-splitter. One can then show that the filter transforms the states in Eq. (5) as follows:

$$\left|\Psi_\ell^-\right\rangle \xrightarrow{\text{HOM filter}} \left|\Psi_\ell^-\right\rangle^{1,2}, \tag{7}$$

$$\left|\Phi_\ell^\pm\right\rangle \xrightarrow{\text{HOM filter}} \frac{i}{2}\left(\left|\Phi_\ell^\pm\right\rangle^{1,1} + \left|\Phi_\ell^\pm\right\rangle^{2,2}\right). \tag{8}$$

Note that the antisymmetric singlet state $\left|\Psi_\ell^-\right\rangle$ exhibits anti-bunching; no two photons are in the same output port. For the two symmetric states, however, photons bunch and exit in the same port of the beam-splitter. Therefore, conditioning the photon detection on coincidence between the two output ports automatically discards the contribution of symmetric states in Eq. (5), leading to the following distillation result:

$$|\Psi_\ell\rangle \xrightarrow{\text{HOM filter}} c_0\left|\Psi_\ell^-\right\rangle^{1,2}. \tag{9}$$

In this manner, noise arising from turbulence is converted to losses, with the probability $|c_0|^2$ representing the fraction of singlets distilled. This fraction naturally depends on the strength of turbulence: with increasing turbulence, $|c_0|$ decays to 0. Given that only anti-symmetric states exhibit anti-bunching after the HOM filter, the choice of the initial singlet state is logical and necessary, that is, because all symmetric states produce no coincidence signal after the HOM filter.[38] Note that the above treatment would also be valid in the case of two photons going through two turbulence screens. This is because given a weak turbulence operator $\hat{M}_i$ acting on photon i we have, within a given OAM subspace,

$$\hat{M}_A \otimes \hat{M}_B\left|\Psi_\ell^-\right\rangle \to \tilde{c_0}\left|\Psi_\ell^-\right\rangle + (\cdots)_{\text{symmetric}}. \tag{10}$$

However, one should expect the fraction of distilled states to be even lower than in the case of one photon going through turbulence, with $|c_0| \geq |\tilde{c_0}|$.

III. EXPERIMENTAL REALISATION

We experimentally demonstrated our distillation scheme using the setup in Fig. 1(a). Pairs of entangled photons were produced through SPDC. We pumped a 3-mm type-I BBO crystal with a 355 nm laser at 350 mW average power and 80 MHz repetition rate, producing the symmetric SPDC state,

$$|\Psi\rangle_{\text{SPDC}} = \alpha_0|0\rangle_A|0\rangle_B + \sum_{\ell>0}\alpha_\ell\left|\Psi_\ell^+\right\rangle_{AB}, \tag{11}$$

where $\left|\Psi_\ell^+\right\rangle_{AB} = (|\ell\rangle_A|-\ell\rangle_B + |-\ell\rangle_A|\ell\rangle_B)/\sqrt{2}$. Using a pair of Dove prisms (DP) in the path of photon A, we controlled the symmetry of OAM subspaces by inducing an OAM-dependent phase in the SPDC state: $|\ell\rangle \to \exp(2i\ell\theta)|\ell\rangle$, where θ is the angle between the two Dove prisms:

$$\left|\Psi_\ell^+\right\rangle \xrightarrow{\text{DPs}} \frac{\exp(2i\ell\theta)|\ell\rangle|-\ell\rangle + \exp(-2i\ell\theta)|-\ell\rangle|\ell\rangle}{\sqrt{2}}. \tag{12}$$

For $\theta = (2n+1)\pi/4\ell$ with $n \in \mathbb{N}$, one achieves the conversion to anti-symmetric state in all the OAM subspaces. We set $\theta = \pi/4$ and obtained, up to a global phase, the following selection rules:

$$\left|\Psi_\ell^+\right\rangle \xrightarrow{\theta=\pi/4} \begin{cases}\left|\Psi_\ell^-\right\rangle \text{ for odd } \ell,\\ \left|\Psi_\ell^+\right\rangle \text{ for even } \ell.\end{cases} \tag{13}$$

Photon A then propagates through turbulence. We considered a turbulent channel with weak scintillation such that atmospheric perturbations can be summed to a unitary transformation, i.e., a single turbulence phase screen,[16] and thus does not lead to a decay in purity. In our case, the turbulence phase screen was modelled based on Kolmogorov's theory and printed on a glass plate with different zones of average phase fluctuations, as shown in Fig. 1(b). In the path of photon B, we placed a delay line, consisting of a Dove prism mounted on a piezo-controlled stage, to adjust the delay in arrival time of the two photons at the BS in the HOM filter. Photons exiting the HOM filter were then analysed using spatial filters encoded on a spatial light modulator (SLM), coupled to single-mode fibres and measured in coincidence.

Initially, we calibrated the HOM filter by scanning for the characteristic dip of coincidence counts in HOM interference. The aim is to demonstrate the efficacy of the distillation scheme by performing a quantum state tomography at points of zero and maximum visibility of the HOM dip; at zero visibility, the HOM filter is in the "OFF" state (out of the HOM interference region), while at maximum visibility, it is in the "ON" state (lowest point in the HOM interference region). Using the digital holograms encoded on the SLM, we post-selected the $\ell = 0$ subspace (symmetric state) from the SPDC state and show the quantum interference signal in Fig. 2(a) in the presence and absence of turbulence. The visibility, $\mathcal{V}$, was computed from the coincidence counts inside (C_{in}) and outside the dip (C_{out}) as follows:

$$\mathcal{V}_{\text{dip}} = \frac{C_{out} - C_{in}}{C_{out} + C_{in}}. \tag{14}$$

We obtained, respectively, a visibility of 84.4% and 75.7% for the dip without and with turbulence. The decay in visibility that we measured can be attributed to the decay in signal-to-noise ratio resulting from turbulence-induced intermodal scattering that reduces the coincidence rates. We will show further that this does affect the distillation of singlet Bell states.

We extended the measurement of the HOM interference trace to the $\ell = \pm 1$ subspace, where the SPDC state has been made antisymmetric. This is done by projecting the photon pair on conjugate OAM states. Due to the anti-bunching effect after the HOM filter, one would expect a HOM peak rather than a dip. This is because events with two-photons in one

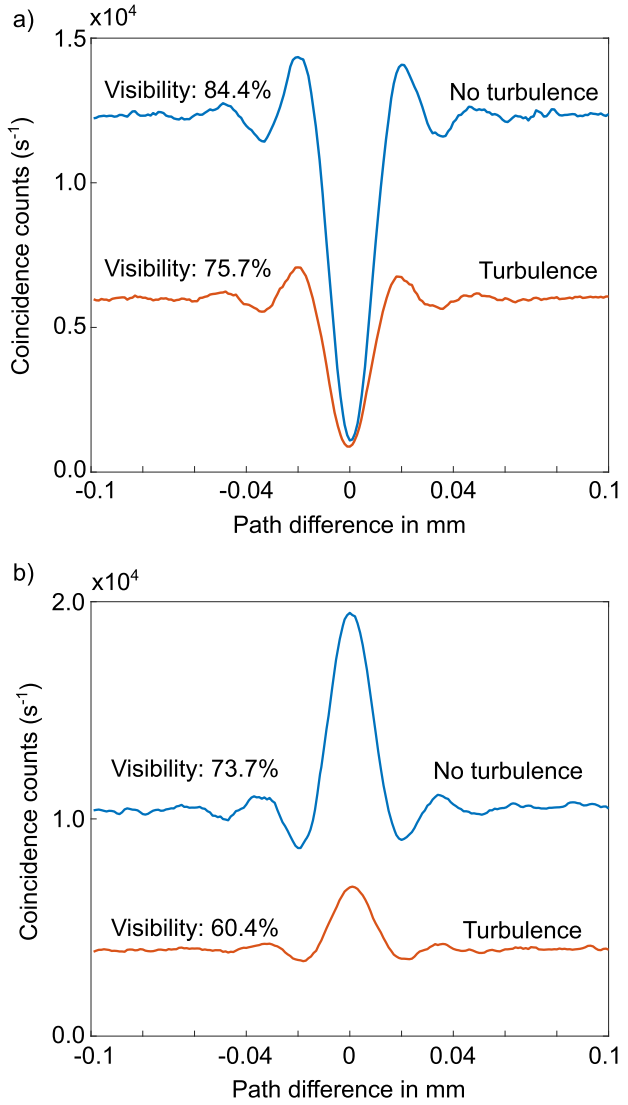


FIG. 2. Hong-Ou-Mandel interference signal for (anti)symmetric states. Delaying one photon with respect to the other leads to (a) a dip and (b) a peak in coincidence counts for the symmetric ($\ell = 0$) and antisymmetric $\ell = \pm 1$ states, respectively. In the presence of turbulence, the visibility of the HOM interference decays in both subspaces, together with the coincidence counts.

output port of the BS are in theory non-existent when the filter is in the ON state. Indeed we measured, as shown in Fig. 2(b), HOM peaks without and with turbulence with visibilities of 73.7% and 60.4%, respectively. In this case, the calculation of visibility was done differently and we will argue the formula we used.

The standard formula in Eq. (14) is adequate for fringes with an ideal minimum of zero; under this condition, the visibility is equal to unity. This is indeed the case for the HOM dip. Due to the bunching effect with the HOM filter ON, symmetric states attain, in principle, a minimum of zero coincidence counts. Therefore the visibility of the dip can be computed as in Eq. (14). For antisymmetric states, coincidence counts should, in theory, increase two-fold when the HOM filter is in the ON state, with no zero minimum. Under Eq. (14), this would lead to negative visibility values with absolute maximum less than unity. Rather, we choose to express the visibility of the HOM peak as follows:

$$\mathcal{V}_{\text{peak}} = \frac{C_{\text{out}} - (2C_{\text{out}} - C_{\text{in}})}{C_{\text{out}} + (2C_{\text{out}} - C_{\text{in}})} = \frac{-C_{\text{out}} + C_{\text{in}}}{3C_{\text{out}} - C_{\text{in}}}. \tag{15}$$

This way of computing the visibility effectively inverts the HOM trace about the minimum (HOM filter OFF), turning the HOM peak into a HOM dip with zero absolute minimum, whose visibility can be calculated as in Eq. (14). Assume a minimum of 1 for the HOM peak signal ($C_{\text{out}} = 1$). Switching the HOM filter to the ON state leads to a maximum coincidence signal of 2 ($C_{\text{in}} = 2$), resulting in a maximum visibility of $\mathcal{V}_{\text{peak}} = 1$.

Having established the reference point for the HOM filter (determination of the dip/peak position), we demonstrated the effectiveness of the distillation process by performing a quantum state tomography of the two-photon state with the HOM filter OFF/ON. Figure 3(a) graphically depicts the joint projective measurements performed within the $\ell = \pm 1$ OAM subspace to realise an over-complete quantum state tomography[40] with the HOM filter in the OFF state. We selected 6 different locations on the surface of the turbulence plate to implement our distillation scheme. The choice of location was solely guided by our need to cover a range of turbulence conditions. With the HOM filter still in the OFF state, we performed a tomography of the two-photon state at the six different locations, as shown in Fig. 3(b). As expected, the measurement outcomes wildly deviate in the presence of turbulence. We then repeated the measurement with the HOM filter in the ON state. Observe that with respect to the reference shown in Fig. 3(c), the tomographic measurements at the previous 6 locations [Fig. 3(d)] are qualitatively similar, indicating that the HOM filter is indeed distilling maximally entangled states.

To each of these tomographic measurements, we can attach a quantitative measure to numerically demonstrate the efficacy of the HOM filter. Our figure of merit here is the fidelity, F, of the reconstructed two-photon density matrix ρ, with respect to the maximally entangled singlet state $\rho_{\text{T}} = |\Psi_\ell^-\rangle\langle\Psi_\ell^-|$,

$$\text{F} = \text{tr}\left(\sqrt{\sqrt{\rho_{\text{T}}}\rho\sqrt{\rho_{\text{T}}}}\right)^2 = \langle\Psi_\ell^-|\rho|\Psi_\ell^-\rangle. \tag{16}$$

For each of the reference measurements in Figs. 3(a) and 3(c), the measured fidelity was in excess of 99%.

When post-selecting the antisymmetric space $\ell = \pm 1$, we show that with the HOM filter OFF, the fidelity with respect to the maximally entangled singlet state decays, as shown in Fig. 4(a), from 0.90 down to 0.095. However, with the HOM filter ON, the fidelity remains constant with F > 0.85 and consistently above the results with the filter OFF. We performed a similar set of measurements with the next antisymmetric subspace $\ell = \pm 3$. States within this subspace are more resilient

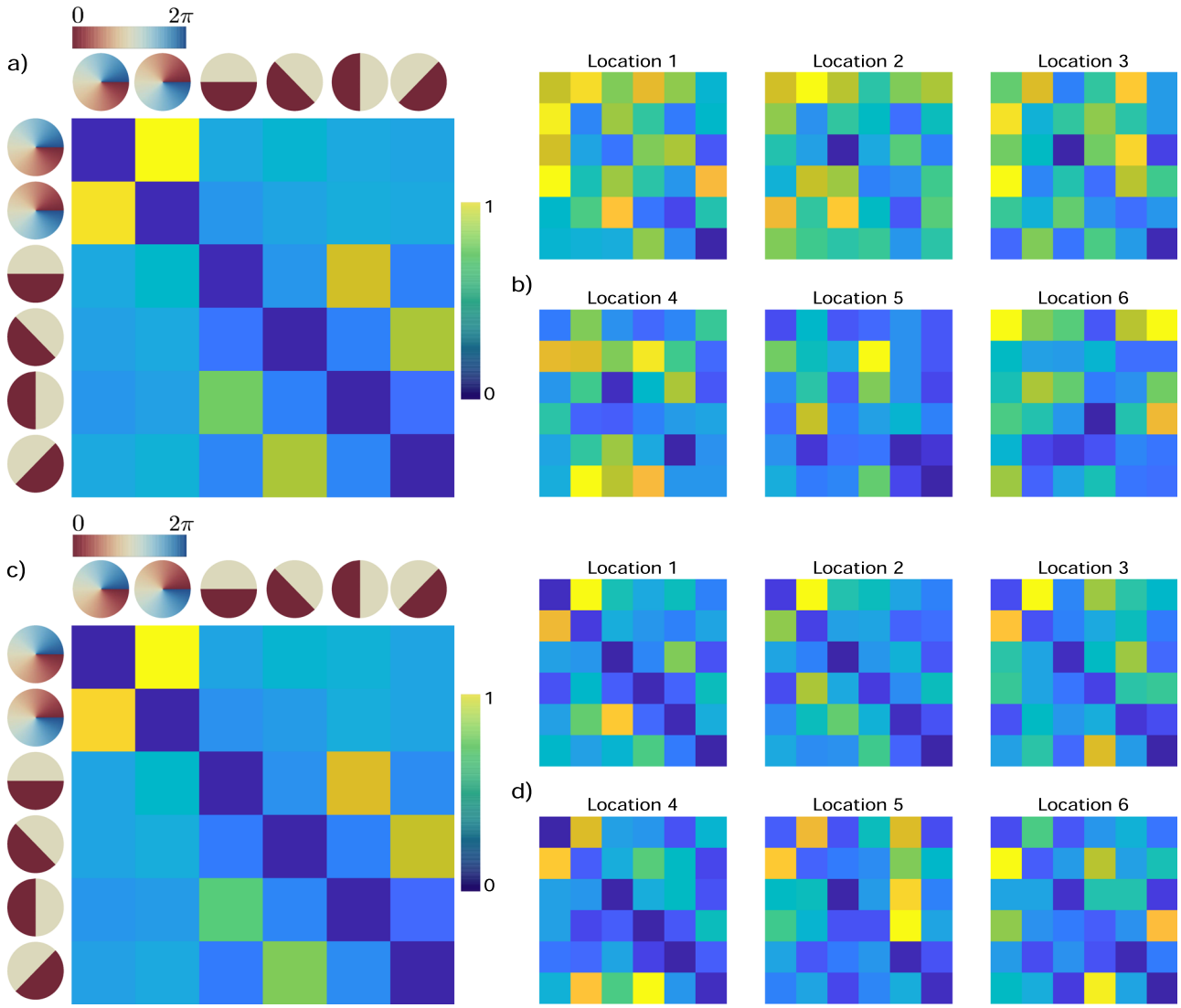


FIG. 3. Quantum state tomography of two-photon states with the HOM filter OFF/ON. (a) shows the normalised tomographic measurement performed with the HOM filter OFF in the absence of turbulence. (b) We introduced the turbulence plate in the path of photon A and show the effect of turbulence on the state tomography measurement outcomes. We repeated the measurements, this time with the HOM filter ON (c) in the absence of turbulence and (d) for the same locations across the turbulence plate.

to turbulence due to their larger separation in OAM space.[23] Using the same locations as before yielded relatively high fidelities without any significant spread. Therefore we chose another set of locations across the turbulence phase plate (TPP) and compared the results with the HOM filter OFF and ON, as shown in Fig. 4(b). Similar to the previous case, with the HOM filter OFF, we measured a decay of fidelity from 0.95 down to 0.032. When switching the filter to the ON state, the fidelity remained consistently high for the various locations, with $F > 0.90$.

The distillation scheme we have demonstrated enables current and future quantum technologies. Turbulence makes it impractical to employ spatial modes for long distance quantum communication, with the measurement fidelity decaying with increasing turbulence. Our scheme enables one to recover, with high fidelity, information that would have otherwise been lost. The resilience of our scheme to a range of turbulence would enable the distribution of photon pairs over significant distances through turbulence. This could be envisaged in a network configuration, where photon pairs are prepared at location A and sent to a remote location B. Upon arrival at location B, photon pairs in the singlet states are distilled, before partaking in other quantum processes that include quantum computation and communication.

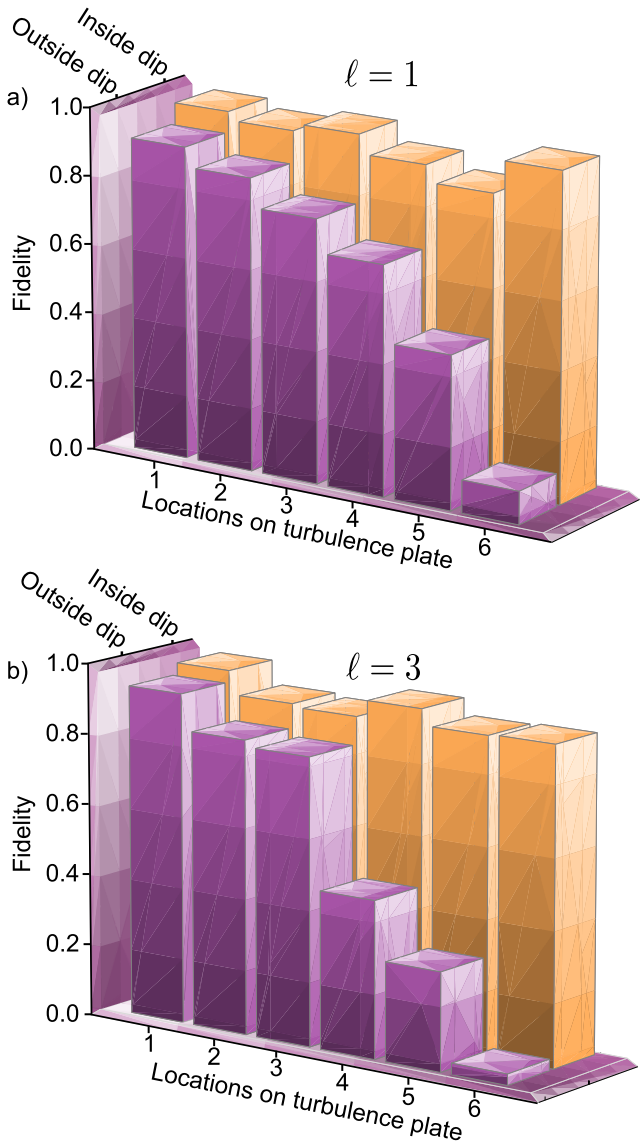


FIG. 4. Distillation of singlet states. With the HOM filter in the OFF state (outside the dip), the fidelity of the measured state with respect to the singlet Bell state decays with increasing perturbations of the turbulence phase plate. Switching the filter to the ON state results in the distillation of states with high-fidelity with respect to the singlet Bell state. This is shown within the (a) $\ell = \pm 1$ and (b) $\ell = \pm 3$ subspaces.

Furthermore, high-fidelity entangled states are critical resources to build a quantum repeater, a fundamental building block of quantum networks.[9] Our distillation scheme can be readily implemented with current technologies.

IV. CONCLUSION

We have demonstrated an entanglement distillation scheme based on quantum interference of two entangled photons. The distillation process is realised by using a Hong-Ou-Mandel filter that causes symmetric and antisymmetric Bell states to respectively bunch and anti-bunch upon exiting the filter through a 50:50 beam-splitter. By conditioning the detection system on coincidence between the output ports of the HOM filter, we have shown theoretically and experimentally the distillation of antisymmetric singlet states. The entanglement concentration process was tested for a coherent superposition of symmetric and antisymmetric states, produced by perturbing a singlet state with a one-sided weak turbulent channel. We have illustrated our distillation scheme by filtering OAM singlets carrying $\ell = \pm\hbar$ and $\ell = \pm 3\hbar$ quanta of OAM from an ensemble of non-maximally entangled pure states. When compared to a singlet Bell state, we have experimentally demonstrated the distillation of states with average fidelity higher than 90% from an ensemble with average fidelity as low as 3%. The states distilled can reliably be used further in other quantum processes.

REFERENCES

[1]S.-K. Liao, W.-Q. Cai, W.-Y. Liu, L. Zhang, Y. Li, J.-G. Ren, J. Yin, Q. Shen, Y. Cao, Z.-P. Li, F.-Z. Li, X.-W. Chen, L.-H. Sun, J.-J. Jia, J.-C. Wu, X.-J. Jiang, J.-F. Wang, Y.-M. Huang, Q. Wang, Y.-L. Zhou, L. Deng, T. Xi, L. Ma, T. Hu, Q. Zhang, Y.-A. Chen, N.-L. Liu, X.-B. Wang, Z.-C. Zhu, C.-Y. Lu, R. Shu, C.-Z. Peng, J.-Y. Wang, and J.-W. Pan, "Satellite-to-ground quantum key distribution," Nature **549**, 43 (2017).

[2]J.-G. Ren, P. Xu, H.-L. Yong, L. Zhang, S.-K. Liao, J. Yin, W.-Y. Liu, W.-Q. Cai, M. Yang, L. Li, K.-X. Yang, X. Han, Y.-Q. Yao, J. Li, H.-Y. Wu, S. Wan, L. Liu, D.-Q. Liu, Y.-W. Kuang, Z.-P. He, P. Shang, C. Guo, R.-H. Zheng, K. Tian, Z.-C. Zhu, N.-L. Liu, C.-Y. Lu, R. Shu, Y.-A. Chen, C.-Z. Peng, J.-Y. Wang, and J.-W. Pan, "Ground-to-satellite quantum teleportation," Nature **549**, 70 (2017).

[3]J. Yin, Y. Cao, Y.-H. Li, S.-K. Liao, L. Zhang, J.-G. Ren, W.-Q. Cai, W.-Y. Liu, B. Li, H. Dai, G.-B. Li, Q.-M. Lu, Y.-H. Gong, Y. Xu, S.-L. Li, F.-Z. Li, Y.-Y. Yin, Z.-Q. Jiang, M. Li, J.-J. Jia, G. Ren, D. He, Y.-L. Zhou, X.-X. Zhang, N. Wang, X. Chang, Z.-C. Zhu, N.-L. Liu, Y.-A. Chen, C.-Y. Lu, R. Shu, C.-Z. Peng, J.-Y. Wang, and J.-W. Pan, "Satellite-based entanglement distribution over 1200 kilometers," Science **356**, 1140 (2017).

[4]V. Giovannetti, S. Lloyd, and L. MacCone, "Advances in quantum metrology," Nat. Photonics **5**, 222 (2011).

[5]V. Giovannetti, S. Lloyd, and L. MacCone, "Quantum-enhanced measurements: Beating the standard quantum limit," Science **306**, 1330 (2004).

[6]P. Kómár, E. M. Kessler, M. Bishof, L. Jiang, A. S. Sørensen, J. Ye, and M. D. Lukin, "A quantum network of clocks," Nat. Phys. **10**, 582 (2014).

[7]J. I. Cirac, A. K. Ekert, S. F. Huelga, and C. Macchiavello, "Distributed quantum computation over noisy channels," Phys. Rev. A **59**, 4249 (1999).

[8]L. Jiang, J. M. Taylor, K. Nemoto, W. J. Munro, R. Van Meter, and M. D. Lukin, "Quantum repeater with encoding," Phys. Rev. A **79**, 032325 (2009).

[9]H. J. Kimble, "The quantum internet," Nature **453**, 1023–1030 (2008).

[10]J. L. O'Brien, "Optical quantum computing," Science **318**, 1567 (2007).

[11]N. Gisin and R. Thew, "Quantum communication," Nat. Photonics **1**, 165 (2007).

[12]A. Mair, A. Vaziri, G. Weihs, and A. Zeilinger, "Entanglement of the orbital angular momentum states of photons," Nature **412**, 313 (2001).

[13]M. Krenn, M. Malik, M. Erhard, and A. Zeilinger, "Orbital angular momentum of photons and the entanglement of Laguerre-Gaussian modes," Philos. Trans. R. Soc., A **375**, 20150442 (2017).

[14]M. Erhard, R. Fickler, M. Krenn, and A. Zeilinger, "Twisted photons: New quantum perspectives in high dimensions," Light: Sci. Appl. **7**, 17146 (2018).

[15]A. C. Dada, J. Leach, G. S. Buller, M. J. Padgett, and E. Andersson, "Experimental high-dimensional two-photon entanglement and violations of generalized Bell inequalities," Nat. Phys. **7**, 677 (2011).

[16]A. Hamadou Ibrahim, F. S. Roux, M. McLaren, T. Konrad, and A. Forbes, "Orbital-angular-momentum entanglement in turbulence," Phys. Rev. A **88**, 012312 (2013).
[17]Y. Zhang, S. Prabhakar, A. H. Ibrahim, F. S. Roux, A. Forbes, and T. Konrad, "Experimentally observed decay of high-dimensional entanglement through turbulence," Phys. Rev. A **94**, 032310 (2016).
[18]C. Gopaul and R. Andrews, "The effect of atmospheric turbulence on entangled orbital angular momentum states," New J. Phys. **9**, 94 (2007).
[19]F. S. Roux, T. Wellens, and V. N. Shatokhin, "Entanglement evolution of twisted photons in strong atmospheric turbulence," Phys. Rev. A **92**, 012326 (2015).
[20]B.-J. Pors, C. H. Monken, E. R. Eliel, and J. P. Woerdman, "Transport of orbital-angular-momentum entanglement through a turbulent atmosphere," Opt. Express **19**, 6671–6683 (2011).
[21]H. Avetisyan and C. H. Monken, "Mode analysis of higher-order transverse-mode correlation beams in a turbulent atmosphere," Opt. Lett. **42**, 101–104 (2017).
[22]S. K. Goyal, A. H. Ibrahim, F. S. Roux, T. Konrad, and A. Forbes, "The effect of turbulence on entanglement-based free-space quantum key distribution with photonic orbital angular momentum," J. Opt. **18**, 064002 (2016).
[23]M. Malik, M. O'Sullivan, B. Rodenburg, M. Mirhosseini, J. Leach, M. P. J. Lavery, M. J. Padgett, and R. W. Boyd, "Influence of atmospheric turbulence on optical communications using orbital angular momentum for encoding," Opt. Express **20**, 13195 (2012).
[24]M. V. da Cunha Pereira, L. A. P. Filpi, and C. H. Monken, "Cancellation of atmospheric turbulence effects in entangled two-photon beams," Phys. Rev. A **88**, 053836 (2013).
[25]C. H. Bennett, H. J. Bernstein, S. Popescu, and B. Schumacher, "Concentrating partial entanglement by local operations," Phys. Rev. A **53**, 2046 (1996).
[26]R. T. Thew and W. J. Munro, "Entanglement manipulation and concentration," Phys. Rev. A **63**, 030302 (2001).
[27]N. A. Peters, J. B. Altepeter, D. Branning, E. R. Jeffrey, T.-C. Wei, and P. G. Kwiat, "Erratum: Maximally entangled mixed states: Creation and concentration [Phys. Rev. Lett. 92, 133601 (2004)]," Phys. Rev. Lett. **96**, 159901 (2006).
[28]P. G. Kwiat, S. Barraza-Lopez, A. Stefanov, and N. Gisin, "Experimental entanglement distillation and 'hidden' non-locality," Nature **409**, 1014 (2001).
[29]S. Bose, V. Vedral, and P. L. Knight, "Purification via entanglement swapping and conserved entanglement," Phys. Rev. A **60**, 194 (1999).
[30]Z. Zhao, J.-W. Pan, and M. S. Zhan, "Practical scheme for entanglement concentration," Phys. Rev. A **64**, 014301 (2001).
[31]Z. Zhao, T. Yang, Y.-A. Chen, A.-N. Zhang, and J.-W. Pan, "Experimental realization of entanglement concentration and a quantum repeater," Phys. Rev. Lett. **90**, 207901 (2003).
[32]T. Yamamoto, M. Koashi, Ş. K. Özdemir, and N. Imoto, "Experimental extraction of an entangled photon pair from two identically decohered pairs," Nature **421**, 343 (2003).
[33]J.-W. Pan, S. Gasparoni, R. Ursin, G. Weihs, and A. Zeilinger, "Experimental entanglement purification of arbitrary unknown states," Nature **423**, 417 (2003).
[34]C. K. Hong, Z. Y. Ou, and L. Mandel, "Measurement of subpicosecond time intervals between two photons by interference," Phys. Rev. Lett. **59**, 2044 (1987).
[35]Y. Zhang, S. Prabhakar, C. Rosales-Guzmán, F. S. Roux, E. Karimi, and A. Forbes, "Hong-Ou-Mandel interference of entangled Hermite-Gauss modes," Phys. Rev. A **94**, 033855 (2016).
[36]Y. Zhang, F. S. Roux, T. Konrad, M. Agnew, J. Leach, and A. Forbes, "Engineering two-photon high-dimensional states through quantum interference," Sci. Adv. **2**, e1501165 (2016).
[37]Y. Zhang, M. Agnew, T. Roger, F. S. Roux, T. Konrad, D. Faccio, J. Leach, and A. Forbes, "Simultaneous entanglement swapping of multiple orbital angular momentum states of light," Nat. Commun. **8**, 632 (2017).
[38]S. Prabhakar, C. Mabena, T. Konrad, and F. S. Roux, "Turbulence and the Hong-Ou-Mandel effect," Phys. Rev. A **97**, 013835 (2018).
[39]M. J. Padgett and J. Courtial, "Poincaré-sphere equivalent for light beams containing orbital angular momentum," Opt. Lett. **24**, 430 (1999).
[40]B. Jack, J. Leach, H. Ritsch, S. M. Barnett, M. J. Padgett, and S. Franke-Arnold, "Precise quantum tomography of photon pairs with entangled orbital angular momentum," New J. Phys. **11**, 103024 (2009).

One-Shot Manipulation of Entanglement for Quantum Channels

Ho-Joon Kim, Soojoon Lee, Ludovico Lami, and Martin B. Plenio

***Abstract*— We show that the dynamic resource theory of quantum entanglement can be formulated using the superchannel theory. In this formulation, we identify the separable channels and the class of free superchannels that preserve channel separability as free resources, and choose the swap channels as dynamic entanglement golden units. Our first result is that the one-shot dynamic entanglement cost of a bipartite quantum channel under the free superchannels is bounded by the standard log-robustness of channels. The one-shot distillable dynamic entanglement of a bipartite quantum channel under the free superchannels is found to be bounded by a resource monotone that we construct from the hypothesis-testing relative entropy of channels with minimization over separable channels. We also address the one-shot catalytic dynamic entanglement cost of a bipartite quantum channel under a larger class of free superchannels that could generate the dynamic entanglement which is asymptotically negligible; it is bounded by the generalized log-robustness of channels.**

***Index Terms*— Quantum entanglement, quantum channel, quantum information theory, quantum resource theory, dynamic resource theory.**

I. Introduction

THE emergence of the modern development of quantum information science is tightly linked to a fundamental change of paradigm that characterizes our appreciation of fundamental traits of quantum mechanics. Rather than viewing these merely as counter-intuitive departures from our classical world view, in recent years we have come to recognize fundamental quantum features as resources that enable us to solve technological and information theoretic tasks more efficiently than classical physics would allow. The desire to investigate systematically which aspects of quantum mechanics are responsible for potential operational advantages has led to the development of quantum resource theories [1]. The most basic concept that gives rise to the structure of resource theories is the concept of constraints that are imposed on our ability to operate beyond those that are already enforced by the laws of quantum mechanics. From this emerges by an elegant inevitability the concept of free states and operations. These are those that can be prepared and executed without violation of the constraints. These two main ingredients allow for the formulation of a rigorous theoretical framework in which to analyze resources quantitatively. Perhaps the most fundamental examples are represented by the theory of quantum coherence, which marks the delineation between classical and quantum physics already at the level of individual particles [2]–[4], and, historically having emerged first, the theory of entanglement, which explores the value of quantum correlations as opposed to classical correlations [5], [6]. These were followed by a host of resource theories including that of superposition [7], [8], of reference frames [9], of Gaussianity [10], of quantum optical non-classicality [11]–[13], of indistinguishable particles [14]–[16], and of thermodynamics [17]–[19].

Initially, the focus of attention in entanglement theory was placed squarely on the entanglement content of quantum states, i.e. (i) which states contain entanglement [20], [21], (ii) how the entanglement of states can be transformed under local operations and classical communication [22]–[25], (iii) how entanglement can be verified quantitatively [26]–[29], and (iv) how useful entanglement is in operational tasks, e.g. to enable the realisation of arbitrary non-local quantum operations when only local operations and classical communication is available. For example, maximally entangled states may be employed to achieve general non-local quantum gates between spatially separated parties using only local quantum operations and classical communication [30], [31]. This example is characteristic of the early approaches to entanglement theory in particular and resource theories in general. While task (ii) concerns state-to-state transformations, task (iv) is of a somewhat different nature, as it connects static resources (states) with dynamic ones (quantum operations).

In fact, quantum states may be considered as special cases of quantum channels, and these subsume also quantum measurements and quantum dynamics [32]–[34]. To make this concrete, consider that quantum states and measurements can be thought of as quantum channels with trivial input and classical output, respectively.

Manuscript received December 21, 2020; accepted April 27, 2021. Date of publication May 13, 2021; date of current version July 14, 2021. This work was supported by the National Research Foundation (NRF) of Korea Grant through the Ministry of Science and ICT (MSIT) under Grant NRF-2019R1A2C1006337 and Grant NRF-2020M3E4A1079678. The work of Soojoon Lee was supported in part by the MSIT through the Information Technology Research Center Support Program supervised by the Institute for Information and Communications Technology Planning and Evaluation under Grant IITP-2021-2018-0-01402 and in part by the Quantum Information Science and Technologies Program of the NRF through the MSIT under Grant NRF-2020M3H3A1105796. The work of Ludovico Lami was supported by the Alexander von Humboldt Foundation. *(Corresponding authors: Ho-Joon Kim; Soojoon Lee.)*

Ho-Joon Kim and Soojoon Lee are with the Department of Mathematics and Research Institute for Basic Sciences, Kyung Hee University, Seoul 02447, South Korea (e-mail: eneration@gmail.com; level@khu.ac.kr).

Ludovico Lami and Martin B. Plenio are with the Institute of Theoretical Physics and IQST, Universität Ulm, 89081 Ulm, Germany (e-mail: ludovico.lami@gmail.com; martin.plenio@uni-ulm.de).

Communicated by S. Beigi, Associate Editor for Quantum Information Theory.

Digital Object Identifier 10.1109/TIT.2021.3079938

While such approach is legitimate based on the fact that any quantum channel can be simulated with free operations and entangled states, the quantum community is now aiming to encompass all aspects in a unified manner, by studying the properties of quantum channels with modern tools of quantum resource theories. First steps in this direction have been taken with the extension of the entropy function from quantum states to quantum channels, and with the identification of its operational meaning in the context of channel merging [33]. Related to this, the resource theory of coherence for quantum channels has been investigated [34]–[37], too. Although notions such as that of relative entropy of entanglement of a quantum channel had been previously utilized in the study of quantum communication [38], this renewed interest in channel resource theories prompted a systematic study of dynamical properties of entanglement [39], resulting in sophisticated theories of dynamic entanglement developed for free resources such as LOCC channels [40], [41] and PPT channels [40]–[44].

In spite of the unifying philosophy underpinning the resource theoretic approach, its extension from the world of states to that of channels is not a trivial task. On the contrary, it presents several challenges that have to do with the fundamentally different nature of the two classes of objects being considered [45]. And yet, the prize of the two undertakings is similar, consisting in the quantitative understanding of numerous operational tasks that are likely to play a prime role in the future of quantum technologies [34]–[36].

In this paper, we will use as basic building blocks specific maximally resourceful operations that play the role that maximally entangled state played in the entanglement theory of states and explore how free superoperations allow us to simulate general quantum channels. As the basic element is itself an operation rather than a state, the nature of the problem justifies the name of dynamic entanglement theory. As a note of caution though, this should not be equated with a theory in which one indeed considers continuous time evolution based on some generators. More specifically, we consider the problem of quantum channel manipulation in the one-shot setting, when the allowed set of free superchannels is maximal, in the sense that it comprises all transformations that map separable channels to separable channels. We choose the K-swap channels as dynamic entanglement resources, which play the role of K-maximally entangled states $\Phi^K_{A_0B_0}$ in the entanglement theory of quantum states. The dynamic entanglement resource is intimately related to the static entanglement resource of quantum states in the sense that the former can generate the maximum static resource under LOCC as well as it requires the maximum static resource to simulate them under LOCC, so it bridges the dynamic entanglement to the maximum static entanglement. In order to treat the dynamic entanglement resource required for conversions between quantum channels, we define the separability-preserving superchannels and use them as the free superchannels. The dynamic entanglement resource of a bipartite quantum channel is investigated in operational ways. To evaluate the one-shot dynamic entanglement cost of a bipartite quantum channel, we ask which amount of dynamic entanglement resource is necessary to simulate the channel by means of free superchannels. To evaluate the one-shot distillable dynamic entanglement of the channel, instead, we determine how much dynamic entanglement resource can be obtained by using only free superchannels. We further push the notion of dynamic entanglement cost to its limits, by considering a two-fold variation on the theme: on the one hand, we allow for catalysts, while on the other we define and use a larger set of free superchannels that might generate a small amount of dynamic entanglement. We refer to the resulting modified notion as the one-shot catalytic dynamic entanglement cost of the channel. Finally, in order to get some insight into the asymptotic scenario, we adopt the liberal smoothing to define liberal dynamic entanglement cost and show that it is given by the liberal regularized relative entropy of channels with respect to the free channels.

II. Dynamic Entanglement Resource

We use indexed capital letters such as A_0, B_1, etc. to denote physical systems, and juxtapose them to indicate physical composites. $\mathcal{B}(\mathcal{H}_{A_0})$ denotes the space of bounded operators acting on a finite dimensional Hilbert space $\mathcal{H}_{A_0}$. The set of linear maps from $\mathcal{B}(\mathcal{H}_{A_0})$ to $\mathcal{B}(\mathcal{H}_{A_1})$ will be denoted with $\mathcal{L}(A) \equiv \mathcal{L}(A_0 \to A_1)$; quantum channels, that is, completely positive trace-preserving linear maps in $\mathcal{L}(A)$, will be collectively denoted with $\mathrm{CPTP}(A) \equiv \mathrm{CPTP}(A_0 \to A_1)$. We use calligraphic letters for quantum channels and abbreviations such as $\mathcal{E}_A \equiv \mathcal{E}_{A_0\to A_1}$. As an exception, we omit indices if we take the trace map over all the input spaces such as $\mathrm{Tr}\,(X_{A_0B_0})$. When a quantum channel applies to a part of a composite system, the identity channel is implicit as in $\mathcal{E}_A(\psi_{A_0B_0}) = \mathcal{E}_A \otimes \mathsf{id}_{B_0}(\psi_{A_0B_0})$. We also omit indices for matrices for readability when there's little chance of confusion. As a distance between two quantum channels, we use the metric induced by the diamond norm denoted as $\|\cdot\|_\diamond$ [46], [47]. $\mathbb{L}(A \to A')$ denotes the set of linear supermaps from $\mathcal{L}(A)$ to $\mathcal{L}(A')$. A Greek letter $\Theta_{A\to B}$ is used to denote supermaps, whose action is expressed as $\Theta_{A\to B}[\mathcal{E}_A]$. We write Ψ_{A_0} for the density matrix of the pure state $|\Psi\rangle_{A_0}$, and call $\mathcal{S}(A_0)$ and $\mathcal{D}(A_0)$ the sets of pure and mixed states of system A_0, respectively. The set of separable states on A_0B_0 is indicated with $\mathrm{SepD}(A_0 : B_0)$, while $\mathrm{SepC}(A : B)$ stands for the set of separable channels from A_0B_0 to A_1B_1.

A K-swap channel $\mathcal{F}^K_{AB}$ consists in the application of the K-swap gate $F^K_{AB} = \sum_{i,j=0}^{K-1} |ij\rangle\langle ji|_{A_0B_0\to A_1B_1}$; it is a typical example of a separability-preserving channel[1] that is not actually separable as a map [49]. We use the K-swap channel $\mathcal{F}^K_{AB}$ as the "golden unit" of resource in the theory of dynamical entanglement [50]; its role is entirely analogous to that of the K-maximally entangled state $|\Phi^K\rangle_{A_0B_0} = \frac{1}{\sqrt{K}} \sum_{i=0}^{K-1} |ii\rangle_{A_0B_0}$ in the static entanglement theory. Indeed, consider that a K-swap channel can generate a pair of K-maximally entangled states between Alice and Bob under LOCC (Fig. 1), while they also need at least two K-maximally entangled states to simulate a K-swap gate with LOCC. Therefore, the golden units for the static and the dynamic entanglement can be regarded equivalent within LOCC.

[1] Also called non-entangling map [48].

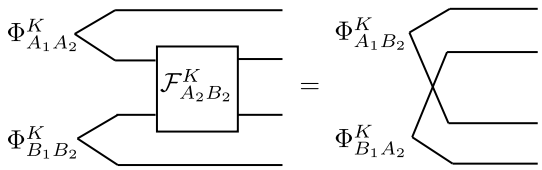


Fig. 1. Two K-maximally entangled states generated by the K-swap channel from locally prepared K-maximally entangled states.

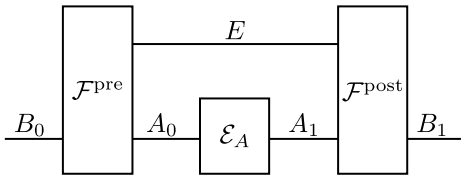


Fig. 2. Structure of a superchannel $\Theta_{A\to B}$ acting on a quantum channel $\mathcal{E}_A$.

A superchannel is a linear supermap that sends a quantum channel to another quantum channel [51], [52]. A superchannel can be decomposed into pre- and post-channel with an additional memory system as in Fig. 2. Taking the separable channels as the free channels, we define the largest set of superchannels that does not generate any dynamic entanglement resource:

Definition II.1: A superchannel $\Theta_{AB\to A'B'}$ is called a *separability-preserving superchannel* (SEPPSC) if

$$\Theta_{AB\to A'B'}[\mathcal{M}_{AB}] \in \text{SepC}(A':B') \tag{1}$$

for all $\mathcal{M}_{AB} \in \text{SepC}(A:B)$. The set of all separability-preserving superchannels from $\mathcal{L}(AB)$ to $\mathcal{L}(A'B')$ is denoted as $\text{SEPPSC}(A:B\to A':B')$.

As dynamic entanglement monotones, we use the standard and the generalized robustness of channels. Since having been introduced for states [53]–[55], these measures have found widespread applications to the quantitative analysis of operational tasks, most notably subchannel discrimination [56]–[58]. The standard robustness of a bipartite quantum channel with respect to the separable channels is defined as

$$R_s(\mathcal{N}_{AB}) \coloneqq \min\left\{ s\geq 0 : \mathcal{M}_{AB}\in \text{SepC}(A:B), \right. \\ \left. \frac{\mathcal{N}_{AB}+s\mathcal{M}_{AB}}{1+s} \in \text{SepC}(A:B)\right\}, \tag{2}$$

while the generalized robustness of a bipartite quantum channel with respect to the separable channels is defined as

$$R(\mathcal{N}_{AB}) \coloneqq \min\left\{ s\geq 0 : \mathcal{M}_{AB}\in \text{CPTP}(AB), \right. \\ \left. \frac{\mathcal{N}_{AB}+s\mathcal{M}_{AB}}{1+s} \in \text{SepC}(A:B)\right\}. \tag{3}$$

The standard log-robustness and the generalized log-robustness of channels with respect to the separable channels are given by

$$LR_s(\mathcal{N}_{AB}) \coloneqq \log\left\{1+R_s(\mathcal{N}_{AB})\right\}, \tag{4}$$
$$LR(\mathcal{N}_{AB}) \coloneqq \log\left\{1+R(\mathcal{N}_{AB})\right\}, \tag{5}$$

respectively, where the logarithm uses base 2. The latter can be expressed as

$$LR(\mathcal{N}_{AB}) = \min_{\mathcal{M}_{AB}\in \text{SepC}(A:B)} D_{\max}(\mathcal{N}_{AB}\|\mathcal{M}_{AB}), \tag{6}$$

where the max-relative entropy between quantum channels is defined as

$$D_{\max}(\mathcal{N}_{AB}\|\mathcal{M}_{AB}) \coloneqq \min\left\{\lambda : \mathcal{N}_{AB}\leq \lambda\mathcal{M}_{AB}\right\}. \tag{7}$$

For $\varepsilon \geq 0$, the smooth versions of the above quantities are defined as

$$LR^{\varepsilon}_s(\mathcal{N}_{AB}) = \min_{\mathcal{N}'_{AB}\approx_{\varepsilon}\mathcal{N}_{AB}} LR_s(\mathcal{N}'_{AB}), \tag{8}$$
$$LR^{\varepsilon}(\mathcal{N}_{AB}) = \min_{\mathcal{N}'_{AB}\approx_{\varepsilon}\mathcal{N}_{AB}} LR(\mathcal{N}'_{AB}), \tag{9}$$

where $\mathcal{N}'_{AB}\approx_{\varepsilon}\mathcal{N}_{AB}$ is a shorthand for $\frac{1}{2}\left\|\mathcal{N}'_{AB}-\mathcal{N}_{AB}\right\|_{\diamond}\leq \varepsilon$. Both robustnesses are monotonically nonincreasing under the free superchannels:

Lemma II.1: For $\Theta_{AB\to A'B'}\in \text{SEPPSC}(A:B\to A':B')$, it holds that

$$R_s(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) \leq R_s(\mathcal{N}_{AB}), \tag{10}$$
$$R(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) \leq R(\mathcal{N}_{AB}). \tag{11}$$

Proof: For $r=R_s(\mathcal{N}_{AB})$, there exist separable channels $\mathcal{M}_{AB}$ and $\mathcal{L}_{AB}$ satisfying that

$$\mathcal{N}_{AB}+r\mathcal{M}_{AB} = (1+r)\mathcal{L}_{AB}. \tag{12}$$

For $\Theta_{AB\to A'B'}\in \text{SEPPSC}(A:B\to A':B')$, we have that

$$\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]+r\Theta_{AB\to A'B'}[\mathcal{M}_{AB}] \\ = (1+r)\Theta_{AB\to A'B'}[\mathcal{L}_{AB}], \tag{13}$$

where $\Theta_{AB\to A'B'}[\mathcal{M}_{AB}]$ and $\Theta_{AB\to A'B'}[\mathcal{L}_{AB}]$ are separable channels. Hence, it follows that $R_s(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}])\leq R_s(\mathcal{N}_{AB})$. The analogous inequality for the generalized robustness follows along the same lines. □

In order to calculate the above quantities for the dynamic entanglement resource, i.e., the K-swap channel, we review previous results on the robustness of bipartite channels, and especially of unitary bipartite channels:

Theorem II.2 [49, Theorem 5]: Let $U_{A_0B_0}$ be a unitary bipartite operator whose operator Schmidt decomposition reads

$$U_{A_0B_0} = \sum_j u_j A_j\otimes B_j, \tag{14}$$

where $A_jA_j^{\dagger}=\frac{I_A}{|A|}$, $B_jB_j^{\dagger}=\frac{I_B}{|B|}$, and $\text{Tr}\,A_j^{\dagger}A_k=\text{Tr}\,B_j^{\dagger}B_k=\delta_{jk}$. Then its robustness is given by

$$R_s(\mathcal{U}_{AB}) = R(\mathcal{U}_{AB}) = \frac{\left(\sum_j u_j\right)^2}{|A|\,|B|}-1. \tag{15}$$

The swap operator F^K_{AB} acting on K-dimensional subsystems can be written as

$$F^K_{AB} = \sum_{i=1}^{K^2} G_i\otimes G_i^{\dagger} \tag{16}$$

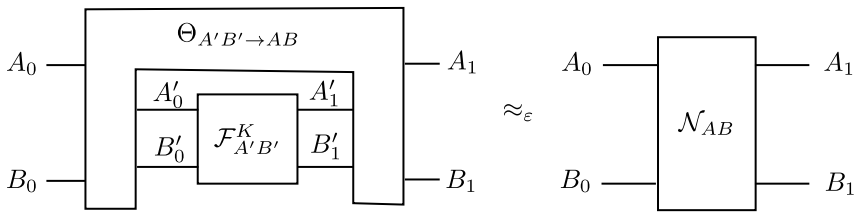


Fig. 3. The one-shot dynamic entanglement cost of a bipartite quantum channel $\mathcal{N}_{AB}$ under a free superchannel $\Theta_{A'B'\to AB}$.

for any orthonormal operator basis $\{G_i\}_{i=1}^{K^2}$ such that $\operatorname{Tr} G_i^\dagger G_j = \delta_{ij}$ [60]. Using an orthonormal unitary basis, e.g., the discrete Weyl basis,[2] the operator Schmidt decomposition of the swap gate is given by

$$F^K_{AB} = \sum_{i=1}^{K^2} \frac{U_i}{\sqrt{K}} \otimes \frac{U_i^\dagger}{\sqrt{K}}. \tag{17}$$

Hence, the robustness of the K-swap channel is given as follows:

$$R_s(\mathcal{F}^K_{AB}) = R(\mathcal{F}^K_{AB}) = K^2 - 1. \tag{18}$$

The following well-known fact concerning separability of the isotropic states will be used afterwards:

Theorem II.3 [61]: The isotropic state

$$p\Phi^K_{A_0B_0} + (1-p)\frac{I_{A_0B_0} - \Phi^K_{A_0B_0}}{K^2-1} \quad (0 \le p \le 1) \tag{19}$$

is separable if and only if $p \le \dfrac{1}{K}$.

III. One-Shot Dynamic Entanglement Cost of a Bipartite Quantum Channel

The first operational task we investigate consists in simulating a single instance of a known channel $\mathcal{N}_{AB}$ using a K-swap channel — with K as small as possible — together with free superchannels, as depicted in Fig. 3. One might call this task dynamic entanglement dilution, in analogy to the entanglement dilution task for quantum states. We can thus give the following formal definition.

Definition III.1: Given $\varepsilon \ge 0$, the one-shot dynamic entanglement cost of a bipartite quantum channel $\mathcal{N}_{AB}$ under SEPPSC is defined as follows:

$$\begin{aligned} E^{(1),\varepsilon}_{C,\text{SEPPSC}}(\mathcal{N}_{AB}) \coloneqq \min\Bigg\{ & \log K^2 : K \in \mathbb{N}_0, \\ & \frac{1}{2}\left\|\Theta_{A'B'\to AB}[\mathcal{F}^K_{A'B'}] - \mathcal{N}_{AB}\right\|_\diamond \le \varepsilon, \\ & \Theta_{A'B'\to AB} \in \text{SEPPSC}(A'\!:\!B' \to A\!:\!B)\Bigg\}. \end{aligned} \tag{20}$$

We now present our first main result. It is a two-fold bound that connects the one-shot dynamic entanglement cost with the smooth standard log-robustness, thus, providing an operational meaning of the latter quantity.

[2]The discrete Weyl basis is composed by the d^2 unitary operators $U_{kl} = \sum_{s=0}^{d-1} e^{\frac{2\pi i}{d}sl}|k+s\rangle\langle s|$, where $k, l = 0, 1, \ldots, d-1$.

Theorem III.1: Given $\varepsilon \ge 0$, the one-shot dynamic entanglement cost of a bipartite quantum channel $\mathcal{N}_{AB}$ under SEPPSC is bounded as

$$LR^\varepsilon_s(\mathcal{N}_{AB}) \le E^{(1),\varepsilon}_{C,\text{SEPPSC}}(\mathcal{N}_{AB}) \le LR^\varepsilon_s(\mathcal{N}_{AB}) + 2. \tag{21}$$

Proof: We break down the argument into separate proofs of the two bounds.

(i) For the lower bound, let

$$\Theta_{A'B'\to AB} \in \text{SEPPSC}(A'\!:\!B' \to A\!:\!B) \tag{22}$$

be a superchannel that achieves $E^{(1),\varepsilon}_{C,\text{SEPPSC}}(\mathcal{N}_{AB})$ with $\mathcal{F}^K_{AB}$, that is, $\Theta_{A'B'\to AB}[\mathcal{F}^K_{A'B'}] \approx_\varepsilon \mathcal{N}_{AB}$. Then we have that

$$\begin{aligned} LR^\varepsilon_s(\mathcal{N}_{AB}) &\le LR_s\left(\Theta_{A'B'\to AB}[\mathcal{F}^K_{A'B'}]\right) \\ &\le LR_s(\mathcal{F}^K_{A'B'}) \\ &= E^{(1),\varepsilon}_{C,\text{SEPPSC}}(\mathcal{N}_{AB}). \end{aligned} \tag{23}$$

(ii) For the upper bound, let $\mathcal{N}^\varepsilon_{AB}$ be a channel such that

$$LR^\varepsilon_s(\mathcal{N}_{AB}) = LR_s(\mathcal{N}^\varepsilon_{AB}) = \log(1+r), \tag{24}$$

where $r = R_s(\mathcal{N}^\varepsilon_{AB})$. There exists a separable channel $\mathcal{M}_{AB}$ such that

$$\frac{\mathcal{N}^\varepsilon_{AB} + r\mathcal{M}_{AB}}{1+r} \in \text{SepC}(A\!:\!B). \tag{25}$$

Let $J^{\mathcal{E}_{AB}}_{A_0B_0\widetilde{A}_1\widetilde{B}_1} \coloneqq \mathsf{id}_{AB} \otimes \mathcal{E}_{\widetilde{A}\widetilde{B}}\left(\Phi^K_{A_0\widetilde{A}_0} \otimes \Phi^K_{B_0\widetilde{B}_0}\right)$ be the (normalized) Choi matrix for a quantum channel $\mathcal{E}_{AB}$, where $|\Phi^K\rangle_{A_0B_0} = \frac{1}{\sqrt{K}}\sum_{i=0}^{K-1}|ii\rangle_{A_0B_0}$ is the maximally entangled state. Setting $K = \lceil\sqrt{1+r}\rceil$, we construct a SEPPSC $\Theta_{A'B'\to AB}$ that simulates $\mathcal{N}^\varepsilon_{AB}$, that is, $\Theta_{A'B'\to AB}[\mathcal{F}^K_{A'B'}] = \mathcal{N}^\varepsilon_{AB}$ as follows:

$$\begin{aligned} \Theta_{A'B'\to AB}[\mathcal{E}_{A'B'}] = {} & \operatorname{Tr}\left(J^{\mathcal{F}^K_{A'B'}} J^{\mathcal{E}_{A'B'}}\right)\mathcal{N}^\varepsilon_{AB} \\ & + \operatorname{Tr}\left\{\left(I - J^{\mathcal{F}^K_{A'B'}}\right) J^{\mathcal{E}_{A'B'}}\right\}\mathcal{M}_{AB}. \end{aligned} \tag{26}$$

While $\Theta_{A'B'\to AB}[\mathcal{F}^K_{A'B'}] = \mathcal{N}^\varepsilon_{AB}$ is apparent from the trace terms, we show that $\Theta_{A'B'\to AB}$ is a SEPPSC. Note that the Choi matrix of a separable channel $\mathcal{E}_{A'B'}$ is a separable state and $J^{\mathcal{F}^K_{A'B'}}_{A_0B_0\widetilde{A}_1\widetilde{B}_1} = \Phi^K_{A_0\widetilde{B}_1} \otimes \Phi^K_{\widetilde{A}_1B_0}$, which leads to

$$\operatorname{Tr}\left(J^{\mathcal{F}^K_{A'B'}} J^{\mathcal{E}_{A'B'}}\right) \le \frac{1}{K^2} \tag{27}$$

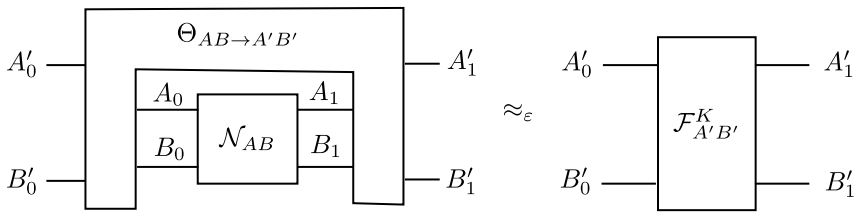


Fig. 4. The one-shot distillable dynamic entanglement of a bipartite quantum channel $\mathcal{N}_{AB}$ under a free superchannel $\Theta_{AB\to A'B'}$.

for any $\mathcal{E}_{A'B'} \in \text{SepC}(A' : B')$.[3] Therefore, when $\mathcal{E}_{A'B'} \in \text{SepC}(A'\!:\!B')$, we have that

$$\begin{aligned}
&\Theta_{A'B'\to AB}[\mathcal{E}_{A'B'}] \\
&= \operatorname{Tr}\left(J^{\mathcal{F}^K_{A'B'}} J^{\mathcal{E}_{A'B'}}\right)\mathcal{N}^\varepsilon_{AB} \\
&\quad + \operatorname{Tr}\left\{\left(I - J^{\mathcal{F}^K_{A'B'}}\right) J^{\mathcal{E}_{A'B'}}\right\}\mathcal{M}_{AB} \\
&= q'\mathcal{N}^\varepsilon_{AB} + (1-q')\mathcal{M}_{AB} \\
&= q\left(\frac{\mathcal{N}^\varepsilon_{AB} + r\mathcal{M}_{AB}}{1+r}\right) + (1-q)\mathcal{M}_{AB} \\
&\in \text{SepC}(A\!:\!B),
\end{aligned} \tag{28}$$

where $q = q'(1+r) \leq 1$ due to $q' \leq \frac{1}{K^2} = \frac{1}{\lceil\sqrt{1+r}\rceil^2}$. We conclude that

$$\begin{aligned}
E^{(1),\varepsilon}_{C,\text{SEPPSC}}(\mathcal{N}_{AB}) &\leq \log K^2 \\
&= 2\log\lceil\sqrt{1+r}\rceil \\
&\leq 2\log(2\sqrt{1+r}) \\
&= \log(1+r) + 2 \\
&\leq LR^\varepsilon_s(\mathcal{N}_{AB}) + 2,
\end{aligned} \tag{29}$$

where in the third line we observed that $\lceil x\rceil \leq 2x$ for all $x \geq 1$. This concludes the proof. □

IV. One-Shot Distillable Dynamic Entanglement of a Bipartite Quantum Channel

The converse task to dynamic entanglement dilution is dynamic entanglement distillation. In our setting, this can be thought of as the task of simulating a K-swap channel — with K as large as possible — using a noisy channel as a dynamic entanglement resource together with free superchannels. We give a pictorial representation of the process in Fig. 4. We can capture this notion through the following formal definition.

Definition IV.1: Given $\varepsilon \geq 0$, the one-shot distillable dynamic entanglement of a bipartite quantum channel $\mathcal{N}_{AB}$ under SEPPSC is defined as

$$\begin{aligned}
E^{(1),\varepsilon}_{D,\text{SEPPSC}}(\mathcal{N}_{AB}) \coloneqq \max\Bigg\{ &\log K^2 : K \in \mathbb{N}_0, \\
&\frac{1}{2}\left\|\Theta_{AB\to A'B'}[\mathcal{N}_{AB}] - \mathcal{F}^K_{A'B'}\right\|_\diamond \leq \varepsilon, \\
&\Theta_{AB\to A'B'} \in \text{SEPPSC}(A\!:\!B \to A'\!:\!B')\Bigg\}.
\end{aligned} \tag{30}$$

[3]This follows directly from a well-known result that is reported in the Appendix as Proposition A.2.

We propose to bound the above operational quantity with a measure that is inspired by the one-shot distillable entanglement of a quantum state [62]. It is obtained from the hypothesis-testing relative entropy of channels by means of an additional minimization over the set of separable channels [63], [64]:

Definition IV.2: Given $\varepsilon \geq 0$, we define the hypothesis-testing relative entropy of dynamic entanglement of a bipartite quantum channel $\mathcal{N}_{AB}$ by

$$\begin{aligned}
E^\varepsilon_H(\mathcal{N}_{AB}) \coloneqq \max_{\Psi}\max_{Q\in S}\min_{\mathcal{M}_{AB}\in\text{SepC}(A:B)} \\
\left\{-\log\operatorname{Tr}\left(Q_{A_1B_1R_0}\cdot\mathcal{M}_{AB}(\Psi_{A_0B_0R_0})\right)\right\},
\end{aligned} \tag{31}$$

where $Q_{A_1B_1R_0}$ is optimized over the set S given by

$$\begin{aligned}
S = \big\{&Q_{A_1B_1R_0} : 0 \leq Q_{A_1B_1R_0} \leq I_{A_1B_1R_0}, \\
&\operatorname{Tr}\left\{Q_{A_1B_1R_0}\cdot\mathcal{N}_{AB}(\Psi_{A_0B_0R_0})\right\} \geq 1-\varepsilon\big\}.
\end{aligned} \tag{32}$$

We remark that the above quantity is monotonic in ε from the definition, implying in particular that $E^\varepsilon_H(\mathcal{N}_{AB}) \geq E^{\varepsilon/2}_H(\mathcal{N}_{AB})$. Moreover, the hypothesis-testing relative entropy of dynamic entanglement does not increase under SEPPSC:

Proposition IV.1: For a bipartite quantum channel $\mathcal{N}_{AB}$, and $\varepsilon \geq 0$, it holds that

$$E^\varepsilon_H(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) \leq E^\varepsilon_H(\mathcal{N}_{AB}) \tag{33}$$

for all $\Theta_{AB\to A'B'} \in \text{SEPPSC}(A\!:\!B \to A'\!:\!B')$.

Proof: Let $\Psi^*_{A'_0B'_0R_0}$ and $Q^*_{A'_1B'_1R_0}$ be optimal arguments of $E^\varepsilon_H(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}])$, so that

$$\begin{aligned}
E^\varepsilon_H(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) = \min_{\mathcal{M}_{A'B'}\in\text{SepC}(A':B')} \\
\left\{-\log\operatorname{Tr} Q^*_{A'_1B'_1R_0}\cdot\mathcal{M}_{A'B'}\left(\Psi^*_{A'_0B'_0R_0}\right)\right\},
\end{aligned} \tag{34}$$

where $0 \leq Q^*_{A'_1B'_1R_0} \leq I_{A'_1B'_1R_0}$ and

$$\begin{aligned}
\operatorname{Tr}\left\{Q^*_{A'_1B'_1R_0}\cdot\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]\left(\Psi^*_{A'_0B'_0R_0}\right)\right\} \\
\geq 1-\varepsilon.
\end{aligned} \tag{35}$$

Then using the structure of the superchannel $\Theta_{AB\to A'B'}[\mathcal{E}_{AB}] = \mathcal{U}_{A_1B_1E_0\to A'_1B'_1}\circ\mathcal{E}_{AB}\circ\mathcal{W}_{A'_0B'_0\to A_0B_0E_0}$ with isometries $\mathcal{U}^\dagger_{A'_1B'_1\to A_1B_1E_0}$ and $\mathcal{W}_{A'_0B'_0\to A_0B_0E_0}$,

we observe that

$$\begin{aligned}
&E_H^{\varepsilon}(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) \\
&= \min_{\mathcal{M}_{A'B'}\in \text{SepC}(A':B')} \Big\{ -\log \operatorname{Tr} Q^*_{A'_1B'_1R_0} \cdot \\
&\qquad \mathcal{M}_{A'B'}\left(\Psi^*_{A'_0B'_0R_0}\right)\Big\} \\
&\leq \min_{\widetilde{\mathcal{M}}_{AB}\in \text{SepC}(A:B)} \Big\{ -\log \operatorname{Tr} Q^*_{A'_1B'_1R_0} \cdot \\
&\qquad \Theta_{AB\to A'B'}\left[\widetilde{\mathcal{M}}_{AB}\right]\left(\Psi^*_{A'_0B'_0R_0}\right)\Big\} \\
&= \min_{\widetilde{\mathcal{M}}_{AB}\in \text{SepC}(A:B)} \Big[-\log \operatorname{Tr}\Big\{ \mathcal{U}^{\dagger}_{A'_1B'_1\to A_1B_1E_0}\left(Q^*_{A'_1B'_1R_0}\right) \cdot \\
&\qquad \widetilde{\mathcal{M}}_{AB}\left(\mathcal{W}_{A'_0B'_0\to A_0B_0E_0}\left(\Psi^*_{A'_0B'_0R_0}\right)\right)\Big\}\Big] \\
&= \min_{\widetilde{\mathcal{M}}_{AB}\in \text{SepC}(A:B)} \Big\{ -\log \operatorname{Tr} \widetilde{Q}^*_{A_1B_1E_0R_0} \cdot \\
&\qquad \widetilde{\mathcal{M}}_{AB}\left(\widetilde{\Psi}^*_{A_0B_0E_0R_0}\right)\Big\} \\
&\leq \max_{\Psi}\max_{Q\in S}\min_{\widetilde{\mathcal{M}}_{AB}\in \text{SepC}(A:B)} \\
&\quad \Big\{ -\log \operatorname{Tr}\left(Q_{A_1B_1E_0R_0}\cdot \widetilde{\mathcal{M}}_{AB}\left(\Psi_{A_0B_0E_0R_0}\right)\right)\Big\} \\
&= E_H^{\varepsilon}(\mathcal{N}_{AB}),
\end{aligned} \tag{36}$$

where the set S is given by

$$\begin{aligned}
S = \big\{ & Q_{A_1B_1E_0R_0} : 0 \leq Q_{A_1B_1E_0R_0} \leq I_{A_1B_1E_0R_0}, \\
& \operatorname{Tr}\left\{Q_{A_1B_1E_0R_0}\cdot \mathcal{N}_{AB}(\Psi_{A_0B_0E_0R_0})\right\} \geq 1-\varepsilon\big\},
\end{aligned} \tag{37}$$

while $\widetilde{Q}^*_{A_1B_1E_0R_0} = \mathcal{U}^{\dagger}_{A'_1B'_1\to A_1B_1E_0}(Q^*_{A'_1B'_1R_0})$ and $\widetilde{\Psi}^*_{A_0B_0E_0R_0} = \mathcal{W}_{A'_0B'_0\to A_0B_0E_0}(\Psi^*_{A'_0B'_0R_0})$. The last inequality holds since $0 \leq \widetilde{Q}^*_{A_1B_1E_0R_0} \leq I_{A_1B_1E_0R_0}$ and

$$\operatorname{Tr}\left\{\widetilde{Q}^*_{A_1B_1E_0R_0}\cdot \mathcal{N}_{AB}\left(\widetilde{\Psi}^*_{A_0B_0E_0R_0}\right)\right\} \geq 1-\varepsilon. \tag{38}$$

This completes the proof. □

Our second main result connects the two notions identified in Definitions IV.1 and IV.2.

Theorem IV.2: Given $\varepsilon \geq 0$ and a bipartite quantum channel $\mathcal{N}_{AB}$, if $\lfloor E_H^{\varepsilon}(\mathcal{N}_{AB})\rfloor$ is even, the one-shot distillable dynamic entanglement from a bipartite quantum channel $\mathcal{N}_{AB}$ under SEPPSC is bounded as

$$\lfloor E_H^{\varepsilon}(\mathcal{N}_{AB})\rfloor \leq E^{(1),\varepsilon}_{D,\text{SEPPSC}}(\mathcal{N}_{AB}) \leq E_H^{2\varepsilon}(\mathcal{N}_{AB}). \tag{39}$$

If $\lfloor E_H^{\varepsilon}(\mathcal{N}_{AB})\rfloor$ is odd, then we have instead that

$$E_H^{\varepsilon}(\mathcal{N}_{AB}) - 1 \leq E^{(1),\varepsilon}_{D,\text{SEPPSC}}(\mathcal{N}_{AB}) \leq E_H^{2\varepsilon}(\mathcal{N}_{AB}). \tag{40}$$

Proof: We break down the argument into separate proofs of the two inequalities.

(i) For the upper bound, let $\Theta_{AB\to A'B'}$ be an optimal SEPPSC such that $\Theta_{AB\to A'B'}[\mathcal{N}_{AB}] \approx_{\varepsilon} \mathcal{F}^K_{A'B'}$ with $E^{(1),\varepsilon}_{D,\text{SEPPSC}}(\mathcal{N}_{AB}) = \log K^2$. From the above two propositions we have that

$$\begin{aligned}
&E_H^{2\varepsilon}(\mathcal{N}_{AB}) \\
&\geq E_H^{2\varepsilon}(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) \\
&= \max_{\Psi}\max_{Q\in S}\min_{\mathcal{M}_{A'B'}\in \text{SepC}(A':B')} \\
&\quad \left[-\log \operatorname{Tr}\left\{Q_{A'_1B'_1R_0}\cdot \mathcal{M}_{A'B'}\left(\Psi_{A'_0B'_0R_0}\right)\right\}\right] \\
&\geq \min_{\mathcal{M}_{A'B'}\in \text{SepC}(A':B')} \\
&\quad \Big\{ -\log \operatorname{Tr} \mathcal{F}^K_{A'B'}\left(\Phi^K_{A'_0\widetilde{A}'_0}\otimes \Phi^K_{B'_0\widetilde{B}'_0}\right)\cdot \\
&\quad \mathcal{M}_{A'B'}\left(\Phi^K_{A'_0\widetilde{A}'_0}\otimes \Phi^K_{B'_0\widetilde{B}'_0}\right)\Big\} \\
&= \log K^2 \\
&= E^{(1),\varepsilon}_{D,\text{SEPPSC}}(\mathcal{N}_{AB}),
\end{aligned} \tag{41}$$

where the set S is given by

$$\begin{aligned}
S = \big\{ & Q_{A'_1B'_1R_0} : 0 \leq Q_{A'_1B'_1R_0} \leq I_{A'_1B'_1R_0}, \\
& \operatorname{Tr}\left\{Q_{A'_1B'_1R_0}\cdot \Theta_{AB\to A'B'}[\mathcal{N}_{AB}]\left(\Psi_{A'_0B'_0R_0}\right)\right\} \\
& \qquad \geq 1-2\varepsilon\big\};
\end{aligned} \tag{42}$$

the second inequality has been derived by making the ansatz $\Psi_{A'_0B'_0R_0} = \Phi^K_{A'_0\widetilde{A}'_0}\otimes \Phi^K_{B'_0\widetilde{B}'_0}$, and the fourth line follows from Proposition A.2. That this is a valid choice is confirmed by the fact that

$$\begin{aligned}
&\frac{1}{2}\left\|\Theta_{AB\to A'B'}[\mathcal{N}_{AB}](\Psi_{A'_0B'_0R_0}) - \mathcal{F}^K_{A'B'}(\Psi_{A'_0B'_0R_0})\right\|_1 \\
&\qquad \leq \frac{1}{2}\left\|\Theta_{AB\to A'B'}[\mathcal{N}_{AB}] - \mathcal{F}^K_{A'B'}\right\|_{\diamond} \leq \varepsilon
\end{aligned} \tag{43}$$

for any (pure) state $\Psi_{A'_0B'_0R_0}$, in turn implying that[4]

$$\begin{aligned}
&F\left(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}](\Psi_{A'_0B'_0R_0}), \mathcal{F}^K_{A'B'}(\Psi_{A'_0B'_0R_0})\right) \\
&\geq \left(1 - \frac{1}{2}\left\|\Theta_{AB\to A'B'}[\mathcal{N}_{AB}](\Psi_{A'_0B'_0R_0})\right.\right. \\
&\qquad \left.\left. - \mathcal{F}^K_{A'B'}(\Psi_{A'_0B'_0R_0})\right\|_1\right)^2 \\
&\geq (1-\varepsilon)^2 \\
&\geq 1-2\varepsilon.
\end{aligned} \tag{44}$$

(ii) For the lower bound, let $\Psi^*_{A_0B_0R_0}$ and $Q^*_{A_1B_1R_0}$ be optimal arguments of $E_H^{\varepsilon}(\mathcal{N}_{AB})$, which satisfy that

$$\begin{aligned}
2^{-E_H^{\varepsilon}(\mathcal{N}_{AB})} = \max_{\mathcal{M}_{AB}\in \text{SepC}(A:B)} & \left\{\operatorname{Tr}\left(Q^*_{A_1B_1R_0}\cdot \right.\right. \\
& \left.\left. \mathcal{M}_{AB}\left(\Psi^*_{A_0B_0R_0}\right)\right)\right\}.
\end{aligned} \tag{45}$$

Setting $K = 2^{\frac{1}{2}\lfloor E_H^{\varepsilon}(\mathcal{N}_{AB})\rfloor}$ for $\lfloor E_H^{\varepsilon}(\mathcal{N}_{AB})\rfloor$ even, and $K = 2^{\left\lfloor \frac{1}{2}E_H^{\varepsilon}(\mathcal{N}_{AB})\right\rfloor}$ otherwise, we can construct a SEPPSC $\Theta_{AB\to A'B'}$ as follows:

$$\begin{aligned}
&\Theta_{AB\to A'B'}[\mathcal{E}_{AB}] \coloneqq \\
&\operatorname{Tr}\left\{Q^*_{A_1B_1R_0}\mathcal{E}_{AB}(\Psi^*_{A_0B_0R_0})\right\}\mathcal{F}^K_{A'B'} \\
&+ \operatorname{Tr}\left\{\left(I_{A_1B_1R_0} - Q^*_{A_1B_1R_0}\right)\mathcal{E}_{AB}(\Psi^*_{A_0B_0R_0})\right\}\mathcal{G}^K_{A'B'},
\end{aligned} \tag{46}$$

[4]Here we are making use of the Fuchs–van de Graaf inequalities [65]. They establish the relations $1-\sqrt{F(\rho,\sigma)} \leq \frac{1}{2}\left\|\rho-\sigma\right\|_1 \leq \sqrt{1-F(\rho,\sigma)}$ between trace distance and quantum fidelity.

where $\mathcal{G}^K_{A'B'}$ is the quantum channel corresponding to the following (normalized) Choi matrix:

$$\begin{aligned} J^{\mathcal{G}^K_{A'B'}} &= \frac{I - J^{\mathcal{F}^K_{A'B'}}}{K^4 - 1} \\ &= \frac{I - \Phi^K_{A'_0\widetilde{B}'_1} \otimes \Phi^K_{\widetilde{A}'_1 B'_0}}{K^4 - 1} \\ &\in \mathrm{SepD}(A'_0\widetilde{A}'_1 : B'_0\widetilde{B}'_1), \end{aligned} \tag{47}$$

which implies that $\mathcal{G}^K_{A'B'} \in \mathrm{SepC}(A' : B')$. For $\mathcal{M}_{AB} \in \mathrm{SepC}(A : B)$, we observe that

$$\begin{aligned} &\Theta_{AB\to A'B'}[\mathcal{M}_{AB}] \\ &:= \mathrm{Tr}\left\{Q^*_{A_1B_1R_0}\mathcal{M}_{AB}(\Psi^*_{A_0B_0R_0})\right\}\mathcal{F}^K_{A'B'} + \\ &\quad \mathrm{Tr}\left\{\left(I_{A_1B_1R_0} - Q^*_{A_1B_1R_0}\right)\mathcal{M}_{AB}(\Psi^*_{A_0B_0R_0})\right\}\mathcal{G}^K_{A'B'} \\ &= q\mathcal{F}^K_{A'B'} + (1-q)\mathcal{G}^K_{A'B'} \\ &\in \mathrm{SepC}(A' : B') \end{aligned} \tag{48}$$

because of $q = \mathrm{Tr}\left\{Q^*_{A_1B_1R_0}\mathcal{M}_{AB}(\Psi^*_{A_0B_0R_0})\right\} \leq \frac{1}{K^2}$ and Theorem II.3 regarding the Choi matrix. Denoting

$$q^* = \mathrm{Tr}\left\{Q^*_{A_1B_1R_0}\mathcal{N}_{AB}(\Psi^*_{A_0B_0R_0})\right\} \geq 1 - \varepsilon, \tag{49}$$

we have that

$$\begin{aligned} &\frac{1}{2}\left\|\Theta_{AB\to A'B'}[\mathcal{N}_{AB}] - \mathcal{F}^K_{A'B'}\right\|_\diamond \\ &= \frac{1}{2}\left\|q^*\mathcal{F}^K_{A'B'} + (1-q^*)\mathcal{G}^K_{A'B'} - \mathcal{F}^K_{A'B'}\right\|_\diamond \\ &\leq \frac{1}{2}\left\|(1-q^*)\mathcal{F}^K_{A'B'}\right\|_\diamond + \frac{1}{2}\left\|(1-q^*)\mathcal{G}^K_{A'B'}\right\|_\diamond \\ &= 1 - q^* \\ &\leq \varepsilon, \end{aligned} \tag{50}$$

where we used that $\|\mathcal{E}\|_\diamond = 1$ for $\mathcal{E} \in \mathrm{CPTP}(AB)$ [66]. Therefore, we conclude that, for $\lfloor E^\varepsilon_H(\mathcal{N}_{AB})\rfloor$ even,

$$E^{(1),\varepsilon}_{D,\mathrm{SEPPSC}}(\mathcal{N}_{AB}) \geq \log K^2 = \lfloor E^\varepsilon_H(\mathcal{N}_{AB})\rfloor. \tag{51}$$

When $\lfloor E^\varepsilon_H(\mathcal{N}_{AB})\rfloor$ is odd, noticing that it holds that $\left\lfloor \frac{q}{p} \right\rfloor \geq \frac{q+1}{p} - 1$ for integers $q, p \in \mathbb{N}$, we obtain that

$$\begin{aligned} E^{(1),\varepsilon}_{D,\mathrm{SEPPSC}}(\mathcal{N}_{AB}) &\geq \log K^2 = 2\left\lfloor \frac{1}{2}E^\varepsilon_H(\mathcal{N}_{AB}) \right\rfloor \\ &\geq 2\left(\frac{E^\varepsilon_H(\mathcal{N}_{AB}) + 1}{2} - 1\right) \\ &= E^\varepsilon_H(\mathcal{N}_{AB}) - 1. \end{aligned} \tag{52}$$

This concludes the proof. □

V. One-Shot Catalytic Dynamic Entanglement Cost of a Bipartite Quantum Channel

The third operational task we consider is a variation on the theme of dynamic entanglement cost. We push this notion further by introducing two tweaks: (i) we allow an additional dynamic entanglement resource that could be used as a catalyst while simulating a bipartite channel, with the stipulation that the catalyst channel be returned intact after the task; and (ii) we introduce a class of superchannels that might generate a small amount of dynamic entanglement when acting on separable channels.

Definition V.1: For $\delta \geq 0$, a superchannel $\Theta_{AB\to A'B'}$ is called δ-separability-preserving superchannel (δ-SEPPSC) if

$$R\left(\Theta_{AB\to A'B'}[\mathcal{M}_{AB}]\right) \leq \delta \quad \forall \mathcal{M}_{AB} \in \mathrm{SepC}(A : B), \tag{53}$$

where R is the generalized robustness with respect to the separable channels.

The choice of the generalized robustness to quantify the maximum amount of entanglement generation allowed in the above definition may seem rather arbitrary. A compelling reason why this is in fact a reasonable and natural choice comes from the study of entanglement theory for states. Indeed, it is known that a condition analogous to (53) leads to a universally reversible theory of entanglement manipulation [48], [67]. The role of the generalized robustness in this context is quite unique, in the sense that using alternative measures — such as the trace norm distance from the set of separable states — is known to trivialize the problem [48, Section V]. In light of this, Definition V.1 identifies a good candidate for a useful enlargement of the set of free superchannels.

As expected, the generalized log-robustness and its smooth version might increase under a δ-separability-preserving superchannel as the following results show:

Proposition V.1: Let $\Theta_{AB\to A'B'}$ be a δ-SEPPSC. For any bipartite quantum channel $\mathcal{N}_{AB}$, the following holds:

$$LR(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) \leq LR(\mathcal{N}_{AB}) + \log(1+\delta). \tag{54}$$

Proof: Let $r \equiv R(\mathcal{N}_{AB})$ such that

$$\mathcal{N}_{AB} + r\mathcal{E}_{AB} = (1+r)\mathcal{M}_{AB}, \tag{55}$$

where $\mathcal{M}_{AB} \in \mathrm{SepC}(A : B)$. It follows that

$$\begin{aligned} &\Theta_{AB\to A'B'}[\mathcal{N}_{AB}] + r\Theta_{AB\to A'B'}[\mathcal{E}_{AB}] \\ &\qquad = (1+r)\Theta_{AB\to A'B'}[\mathcal{M}_{AB}]. \end{aligned} \tag{56}$$

Also, we have that

$$\Theta_{AB\to A'B'}[\mathcal{M}_{AB}] + r'\mathcal{G}_{A'B'} = (1+r')\mathcal{M}'_{A'B'}, \tag{57}$$

where $r' \equiv R(\Theta_{AB\to A'B'}[\mathcal{M}_{AB}]) \leq \delta$, and $\mathcal{M}'_{A'B'} \in \mathrm{SepC}(A' : B')$. From these two equations, it follows that

$$\begin{aligned} &\Theta_{AB\to A'B'}[\mathcal{N}_{AB}] + r\Theta_{AB\to A'B'}[\mathcal{E}_{AB}] + (1+r)r'\mathcal{G}_{A'B'} \\ &\qquad = (1+r)(1+r')\mathcal{M}'_{A'B'}, \end{aligned} \tag{58}$$

which implies that

$$1 + R(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) \leq (1+r)(1+r'). \tag{59}$$

□

Lemma V.2: For $\Theta_{AB\to A'B'} \in \delta\text{-SEPPSC}(A : B \to A' : B')$, we have that

$$LR^\varepsilon(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) \leq LR^\varepsilon(\mathcal{N}_{AB}) + \log(1+\delta). \tag{60}$$

Proof: Let $\mathcal{N}^\varepsilon_{AB}$ be a quantum channel satisfying that $LR^\varepsilon(\mathcal{N}_{AB}) = LR(\mathcal{N}^\varepsilon_{AB})$. We have that

$$\begin{aligned} LR^\varepsilon(\Theta_{AB\to A'B'}[\mathcal{N}_{AB}]) &\leq LR(\Theta_{AB\to A'B'}[\mathcal{N}^\varepsilon_{AB}]) \\ &\leq LR(\mathcal{N}^\varepsilon_{AB}) + \log(1+\delta) \\ &= LR^\varepsilon(\mathcal{N}_{AB}) + \log(1+\delta), \end{aligned}$$

concluding the proof. □

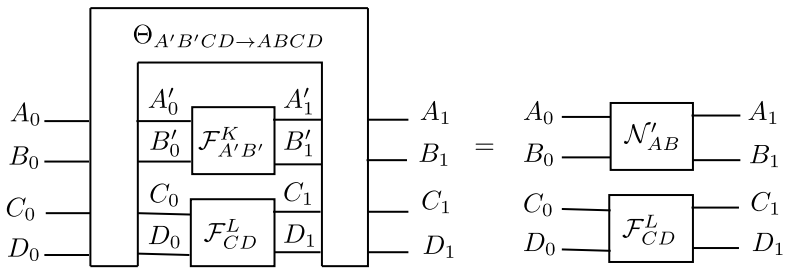


Fig. 5. One-shot catalytic dynamic entanglement cost of a bipartite quantum channel $\mathcal{N}_{AB}$ under δ-SEPPSC $\Theta_{A'B'CD\to ABCD}$, where $\mathcal{N}'_{AB}\approx_\varepsilon \mathcal{N}_{AB}$.

With these tools, we give the formal definition of the one-shot catalytic dynamic entanglement cost of a bipartite channel as follows:

Definition V.2: Given $\delta > 0$ and $\varepsilon \geq 0$, the one-shot catalytic dynamic entanglement cost of a bipartite quantum channel $\mathcal{N}_{AB}$ under δ-SEPPSC is defined as

$$\begin{aligned}\widetilde{E}^{(1),\varepsilon}_{C,\delta\text{-SEPPSC}}(\mathcal{N}_{AB}) \coloneqq \min\Big\{ &\log K^2 : K, L \in \mathbb{N}_0,\\ &\Theta_{A'B'CD\to ABCD}[\mathcal{F}^K_{A'B'}\otimes\mathcal{F}^L_{CD}] = \mathcal{N}'_{AB}\otimes\mathcal{F}^L_{CD},\\ &\Theta_{A'B'CD\to ABCD}\in \delta\text{-SEPPSC}(A'C:B'D\to AC:BD),\\ &\frac{1}{2}\left\|\mathcal{N}'_{AB}-\mathcal{N}_{AB}\right\|_\diamond \leq \varepsilon\Big\}. \end{aligned} \tag{61}$$

We depict the operational task in Fig. 5. In order to bound the one-shot catalytic dynamic entanglement cost of a bipartite channel, the following lemma uses a twisted twirling superchannel that separates the K-swap channel from the others.

Lemma V.3: For a bipartite quantum channel $\mathcal{N}_{AB}$ and $\varepsilon \geq 0$, there is a quantum channel $\mathcal{M}^\varepsilon_{ABCD}$ given by

$$\mathcal{M}^\varepsilon_{ABCD} = p\mathcal{N}^\varepsilon_{AB}\otimes\mathcal{F}^L_{CD} + (1-p)\mathcal{L}_{ABCD}, \tag{62}$$

where $\mathcal{L}_{ABCD}$ is a quantum channel, $\frac{1}{2}\left\|\mathcal{N}^\varepsilon_{AB}-\mathcal{N}_{AB}\right\|_\diamond \leq |A_0|\,|B_0|\sqrt{2\varepsilon}$, and $p\geq 1-2\varepsilon$. It also satisfies that

$$LR(\mathcal{M}^\varepsilon_{ABCD}) \leq LR^\varepsilon(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}). \tag{63}$$

Proof: Let $\widetilde{\mathcal{M}}^\varepsilon_{ABCD}$ be a quantum channel satisfying

$$LR^\varepsilon(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}) = LR(\widetilde{\mathcal{M}}^\varepsilon_{ABCD}) \equiv l, \tag{64}$$

which implies the existence of a separable channel $\Sigma_{ABCD}\in \mathrm{SepC}(AC:BD)$ such that

$$\widetilde{\mathcal{M}}^\varepsilon_{ABCD} \leq 2^l\Sigma_{ABCD}. \tag{65}$$

Since $\widetilde{\mathcal{M}}^\varepsilon_{ABCD}$ is ε-close to $\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}$ by definition, we expect to have $\mathcal{M}^\varepsilon_{ABCD}$ by properly pinching it. We try the following twisted twirling superchannel, which can be performed via LOCC:

$$\Omega_{AB}[\mathcal{E}_{AB}] \coloneqq \iint \mathcal{U}_{A_1}\otimes\mathcal{V}_{B_1}\circ\mathcal{E}_{AB}\circ\mathcal{V}^\dagger_{A_0}\otimes\mathcal{U}^\dagger_{B_0}. \tag{66}$$

For a quantum channel $\mathcal{E}_{AB}$, the twisted twirling superchannel turns its (normalized) Choi matrix into a structured form:

$$\begin{aligned} J^{\Omega_{AB}[\mathcal{E}_{AB}]} &= \iint \overline{\mathcal{V}}_{A_0}\otimes\overline{\mathcal{U}}_{B_0}\otimes\mathcal{U}_{\widetilde{A}_1}\otimes\mathcal{V}_{\widetilde{B}_1}\left(J^{\mathcal{E}_{AB}}\right)\\ &= p_0\,\Phi^K_{A_0\widetilde{B}_1}\otimes\Phi^K_{\widetilde{A}_1B_0} + p_1\,\Phi^K_{A_0\widetilde{B}_1}\otimes\frac{I-\Phi^K_{\widetilde{A}_1B_0}}{K^2-1}\\ &\quad + p_2\,\frac{I-\Phi^K_{A_0\widetilde{B}_1}}{K^2-1}\otimes\Phi^K_{\widetilde{A}_1B_0}\\ &\quad + p_3\,\frac{I-\Phi^K_{A_0\widetilde{B}_1}}{K^2-1}\otimes\frac{I-\Phi^K_{\widetilde{A}_1B_0}}{K^2-1}\\ &= p_0J^{\mathcal{F}^K_{AB}} + (1-p_0)J^{\mathcal{Q}_{AB}}. \end{aligned}$$

Note that $\mathrm{Tr}\left(J^{\mathcal{F}^K_{AB}}J^{\mathcal{Q}_{AB}}\right) = 0$. Applying the twisted twirling superchannel Ω_{CD} on $\widetilde{\mathcal{M}}^\varepsilon_{ABCD}$, we devise $\mathcal{M}^\varepsilon_{ABCD}$ by as follows:

$$\begin{aligned}\mathcal{M}^\varepsilon_{ABCD} &= \Omega_{CD}\left[\widetilde{\mathcal{M}}^\varepsilon_{ABCD}\right]\\ &= p\mathcal{N}^\varepsilon_{AB}\otimes\mathcal{F}^L_{CD} + (1-p)\mathcal{L}_{ABCD}.\end{aligned}$$

We show that $\mathcal{M}^\varepsilon_{ABCD}$ satisfies the insisted properties. Firstly, by construction,

$$\mathcal{M}^\varepsilon_{ABCD} = \Omega_{CD}\left[\widetilde{\mathcal{M}}^\varepsilon_{ABCD}\right] \leq 2^l\Omega_{CD}\left[\Sigma_{ABCD}\right]. \tag{67}$$

Since Ω_{CD} can be done by LOCC, we have $\Omega_{CD}\left[\Sigma_{ABCD}\right]\in \mathrm{SepC}(AC:BD)$. Therefore, we have

$$LR(\mathcal{M}^\varepsilon_{ABCD}) \leq LR^\varepsilon(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}). \tag{68}$$

From the contractivity of the diamond distance under a superchannel, it follows that

$$\begin{aligned}\varepsilon &\geq \frac{1}{2}\left\|\widetilde{\mathcal{M}}^\varepsilon_{ABCD}-\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}\right\|_\diamond\\ &\geq \frac{1}{2}\left\|\Omega_{CD}\left[\widetilde{\mathcal{M}}^\varepsilon_{ABCD}\right]-\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}\right\|_\diamond,\end{aligned}$$

where we used $\Omega_{CD}[\mathcal{F}^L_{CD}] = \mathcal{F}^L_{CD}$. Using Theorem A.4, we get to

$$\begin{aligned}1-2\varepsilon &\leq F\left(J^{\Omega_{CD}[\widetilde{\mathcal{M}}^\varepsilon_{ABCD}]}, J^{\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}}\right)\\ &= F\left(pJ^{\mathcal{N}^\varepsilon_{AB}\otimes\mathcal{F}^L_{CD}} + (1-p)J^{\mathcal{L}_{ABCD}}, J^{\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}}\right)\\ &= pF\left(J^{\mathcal{N}^\varepsilon_{AB}\otimes\mathcal{F}^L_{CD}}, J^{\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}}\right)\\ &= pF\left(J^{\mathcal{N}^\varepsilon_{AB}}, J^{\mathcal{N}_{AB}}\right),\end{aligned}$$

where the second equality follows from the orthogonality of the Choi matrices. From the above, we read that $p \geq 1-2\varepsilon$ and $F\left(J^{\mathcal{N}^{\varepsilon}_{AB}}, J^{\mathcal{N}_{AB}}\right) \geq 1-2\varepsilon$ due to $p \leq 1$ and $F\left(J^{\mathcal{N}^{\varepsilon}_{AB}}, J^{\mathcal{N}_{AB}}\right) \leq 1$. Furthermore, $F\left(J^{\mathcal{N}^{\varepsilon}_{AB}}, J^{\mathcal{N}_{AB}}\right) \geq 1-2\varepsilon$ together with Theorem A.3 and the Fuchs-van der Graaf inequality implies the following:

$$\begin{aligned}\frac{1}{2}\left\|\mathcal{N}^{\varepsilon}_{AB}-\mathcal{N}_{AB}\right\|_{\diamond} &\leq |A_0|\,|B_0|\,\frac{1}{2}\left\|J^{\mathcal{N}^{\varepsilon}_{AB}}-J^{\mathcal{N}_{AB}}\right\|_1 \\ &\leq |A_0|\,|B_0|\sqrt{1-F\left(J^{\mathcal{N}^{\varepsilon}_{AB}}, J^{\mathcal{N}_{AB}}\right)} \\ &\leq |A_0|\,|B_0|\sqrt{2\varepsilon}.\end{aligned}$$

This completes the proof. □

We bound the one-shot catalytic dynamic entanglement cost of a bipartite channel as follows:

Theorem V.4: Given $\delta > 0$, $\varepsilon \geq 0$, there exists $L \in \mathbb{N}$ such that $L^2 \geq 1+\frac{1}{\delta}$, and the one-shot catalytic dynamic entanglement cost for any bipartite quantum channel $\mathcal{N}_{AB}$ is bounded as

$$\begin{aligned}&LR^{\varepsilon}(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}) - \log L^2 - \log(1+\delta) \\ &\quad\leq \widetilde{E}^{(1),\varepsilon}_{C,\delta\text{-SEPPSC}}(\mathcal{N}_{AB}) \\ &\quad\leq LR^{\varepsilon'}(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}) - \log L^2 - \log(1-2\varepsilon') + 2,\end{aligned}$$

where $\varepsilon' = \varepsilon^2/\left(2\,|A_0|^2\,|B_0|^2\right)$.

Proof: We break down the argument into separate proofs of the two bounds.

(i) For the upper bound, let $\mathcal{M}^{\varepsilon}_{ABCD}$ be a quantum channel as in Lemma V.3 satisfying that

$$\begin{aligned}LR(\mathcal{M}^{\varepsilon'}_{ABCD}) &\leq LR^{\varepsilon'}(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD}), \\ \mathcal{M}^{\varepsilon'}_{ABCD} &= p\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} + (1-p)\mathcal{L}_{ABCD}, \\ \frac{1}{2}\left\|\mathcal{N}^{\varepsilon}_{AB}-\mathcal{N}_{AB}\right\|_{\diamond} &\leq \varepsilon, \\ p &\geq 1-2\varepsilon',\end{aligned}$$

where $\varepsilon' = \varepsilon^2/\left(2\,|A_0|^2\,|B_0|^2\right)$. From the first and the second equation above, it follows that

$$\begin{aligned}\mathcal{M}^{\varepsilon'}_{ABCD} &= p\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} + (1-p)\mathcal{L}_{ABCD} \\ &\leq 2^{LR^{\varepsilon'}(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD})}\Sigma_{ABCD},\end{aligned}$$

where $\Sigma_{ABCD} \in \text{SepC}(AC:BD)$. Since $\mathcal{L}_{ABCD} \geq 0$, we have that

$$\begin{aligned}\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} &\leq 2^{LR^{\varepsilon'}(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD})-\log p}\Sigma_{ABCD} \\ &\leq 2^{LR^{\varepsilon'}(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD})-\log(1-2\varepsilon')}\Sigma_{ABCD},\end{aligned}$$

which leads to the existence of a quantum channel $\mathcal{R}_{ABCD}$ such that

$$\frac{\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} + (r-1)\mathcal{R}_{ABCD}}{r} \in \text{SepC}(AC:BD), \tag{69}$$

where we denote $r = 2^{LR^{\varepsilon'}(\mathcal{N}_{AB}\otimes\mathcal{F}^L_{CD})-\log(1-2\varepsilon')}$. With the insight gained above, we construct a superchannel $\Theta_{A'B'CD\to ABCD} \in \delta\text{-SEPPSC}(A'C:B'D\to AC:BD)$ as follows:

$$\begin{aligned}&\Theta_{A'B'CD\to ABCD}[\mathcal{E}_{A'B'CD}] \\ &:= \text{Tr}\left(J^{\mathcal{F}^K_{A'B'}\otimes\mathcal{F}^L_{CD}}J^{\mathcal{E}_{A'B'CD}}\right)\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} \\ &\quad+\text{Tr}\left\{\left(I-J^{\mathcal{F}^K_{A'B'}\otimes\mathcal{F}^L_{CD}}\right)J^{\mathcal{E}_{A'B'CD}}\right\}\mathcal{R}_{ABCD},\end{aligned} \tag{70}$$

where we use the (normalized) Choi matrix and the identity matrix omitting some of the indices for brevity. It is obvious that

$$\Theta_{A'B'CD\to ABCD}[\mathcal{F}^K_{A'B'}\otimes\mathcal{F}^L_{CD}] = \mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} \tag{71}$$

from the construction. We now show that $\Theta_{A'B'CD\to ABCD} \in \delta\text{-SEPPSC}(A'C:B'D\to AC:BD)$. For $\mathcal{E}_{A'B'CD} \in \text{SepC}(A'C:B'D)$, we have that

$$\begin{aligned}&\Theta_{A'B'CD\to ABCD}[\mathcal{E}_{A'B'CD}] \\ &= \text{Tr}\left(J^{\mathcal{F}^K_{A'B'}\otimes\mathcal{F}^L_{CD}}J^{\mathcal{E}_{A'B'CD}}\right)\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} \\ &\quad+\text{Tr}\left\{\left(I-J^{\mathcal{F}^K_{A'B'}\otimes\mathcal{F}^L_{CD}}\right)J^{\mathcal{E}_{A'B'CD}}\right\}\mathcal{R}_{ABCD} \\ &= q\,\frac{\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} + (r-1)\mathcal{R}_{ABCD}}{r} + (1-q)\mathcal{R}_{ABCD},\end{aligned} \tag{72}$$

where $q = r\,\text{Tr}\left(J^{\mathcal{F}^K_{A'B'}\otimes\mathcal{F}^L_{CD}}J^{\mathcal{E}_{A'B'CD}}\right) \leq \frac{r}{K^2L^2}$ due to Proposition A.2. Setting $K = \left\lceil\frac{\sqrt{r}}{L}\right\rceil$, it is assured that $q \leq 1$. Therefore, from the convexity of the robustness, it follows that

$$\begin{aligned}&R\left(\Theta_{A'B'CD\to ABCD}[\mathcal{E}_{A'B'CD}]\right) \\ &\leq\ q\,R\left(\frac{\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD} + (r-1)\mathcal{R}_{ABCD}}{r}\right) \\ &\quad+(1-q)R\left(\mathcal{R}_{ABCD}\right) \\ &\leq\ R\left(\mathcal{R}_{ABCD}\right).\end{aligned} \tag{73}$$

Furthermore, we have that

$$\begin{aligned}R\left(\mathcal{R}_{ABCD}\right) &\leq \frac{1}{R\left(\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD}\right)} \\ &\leq \frac{1}{R\left(\mathcal{F}^L_{CD}\right)} \\ &= \frac{1}{L^2-1},\end{aligned} \tag{74}$$

where the first inequality follows from equation (69), and the second inequality follows from the monotonicity of the robustness under SEPPSC,[5] that is, $R\left(\mathcal{N}^{\varepsilon}_{AB}\otimes\mathcal{F}^L_{CD}\right) \geq R\left(\mathcal{F}^L_{CD}\right)$. Thus, if $R\left(\mathcal{R}_{ABCD}\right) \leq \frac{1}{L^2-1} \leq \delta$, or $L^2 \geq 1+\frac{1}{\delta}$, then for $\mathcal{E}_{A'B'CD} \in \text{SepC}(A'C:B'D)$, we have that

$$R\left(\Theta_{A'B'CD\to ABCD}[\mathcal{E}_{A'B'CD}]\right) \leq R\left(\mathcal{R}_{ABCD}\right) \leq \delta. \tag{75}$$

[5]One can feed any product state $\rho_A\otimes\rho_B$ into the quantum channel and subsequently trace away some subsystems at the output.

So we have that $\Theta_{A'B'CD\to ABCD} \in \delta\text{-SEPPSC}(A'C : B'D \to AC : BD)$ by setting K as

$$K = \left\lceil \frac{\sqrt{r}}{L} \right\rceil = \left\lceil \frac{\sqrt{2^{LR^{\varepsilon'}(\mathcal{N}_{AB}\otimes\mathcal{F}_{CD}^{L})-\log(1-2\varepsilon')}}}{L} \right\rceil, \tag{76}$$

where $L \in \mathbb{N}$ is chosen to satisfy that $L^2 \geq 1 + \frac{1}{\delta}$. Finally, we conclude that

$$\begin{aligned} &\widetilde{E}^{(1),\varepsilon}_{C,\delta\text{-SEPPSC}}(\mathcal{N}_{AB}) \\ &\leq \log K^2 \\ &\leq LR^{\varepsilon'}\left(\mathcal{N}_{AB}\otimes\mathcal{F}_{CD}^{L}\right) - \log L^2 - \log(1-2\varepsilon') + 2, \end{aligned} \tag{77}$$

where $\varepsilon' = \varepsilon^2 / \left(2\left|A_0\right|^2\left|B_0\right|^2\right)$.

(ii) For the lower bound, let $\widetilde{E}^{(1),\varepsilon}_{C,\delta\text{-SEPPSC}}(\mathcal{N}_{AB}) = \log K^2$ for which a catalyst $\mathcal{F}_{CD}^{L_0}$ is used as follows:

$$\Theta_{A'B'CD\to ABCD}[\mathcal{F}_{A'B'}^{K}\otimes\mathcal{F}_{CD}^{L_0}] = \mathcal{N}'_{AB}\otimes\mathcal{F}_{CD}^{L_0}, \tag{78}$$

$$\mathcal{N}'_{AB} \approx_{\varepsilon} \mathcal{N}_{AB}, \tag{79}$$

where $\Theta_{A'B'CD\to ABCD} \in \delta\text{-SEPPSC}(A'C : B'D \to AC : BD)$. It follows that

$$\begin{aligned} &LR^{\varepsilon}\left(\mathcal{N}_{AB}\otimes\mathcal{F}_{CD}^{L_0}\right) \\ &\leq LR\left(\mathcal{N}'_{AB}\otimes\mathcal{F}_{CD}^{L_0}\right) \\ &= LR\left(\Theta_{A'B'CD\to ABCD}[\mathcal{F}_{A'B'}^{K}\otimes\mathcal{F}_{CD}^{L_0}]\right) \\ &\leq LR\left(\mathcal{F}_{A'B'}^{K}\otimes\mathcal{F}_{CD}^{L_0}\right) + \log(1+\delta) \\ &= \log K^2 + \log L_0^2 + \log(1+\delta). \end{aligned} \tag{80}$$

Moreover, we can choose a universal lower bound $\widetilde{L}_0$ that satisfies the following:

$$\begin{aligned} &LR^{\varepsilon}(\mathcal{N}_{AB}\otimes\mathcal{F}_{CD}^{\widetilde{L}_0}) - \log\widetilde{L}_0^2 - \log(1+\delta) \\ &\quad\leq \widetilde{E}^{(1),\varepsilon}_{C,\delta\text{-SEPPSC}}(\mathcal{N}_{AB}) \quad \forall\mathcal{N}_{AB}\in\text{CPTP}(AB). \end{aligned} \tag{81}$$

Finally, regarding (i) and (ii), we can choose $L = \max\left\{\widetilde{L}_0, \left\lceil\sqrt{1+\frac{1}{\delta}}\right\rceil\right\}$ which provides both the upper and the lower bound on $\widetilde{E}^{(1),\varepsilon}_{C,\delta\text{-SEPPSC}}(\mathcal{N}_{AB})$ for any bipartite quantum channel $\mathcal{N}_{AB}$). This completes the proof. □

VI. Conclusion

We found that entanglement of quantum channels can be naturally understood adopting the superchannel framework. Taking the separable channels as our free resource, we defined the separability-preserving superchannels as the resource non-generating superchannels. The K-swap channel $\mathcal{F}_{AB}^{K}$ is chosen as the dynamic entanglement resource, mimicking the role of the K-maximally entangled state in the resource theory of static entanglement. In fact, these two objects are totally interchangeable, because a K-swap channel can be transformed into a pair of K-maximally entangled states under LOCC, and vice versa, at least two K-maximally entangled states are necessary to implement a K-swap channel with LOCC — more precisely, by performing two times a teleportation protocol. Our results provide an operational meaning to the standard and the generalized log-robustness of channels as well as the hypothesis-testing relative entropy of dynamic entanglement that we constructed from the hypothesis-testing relative entropy of channels with minimization over the set of separable channels: The one-shot dynamic entanglement cost can be bounded by the standard log-robustness of channels with respect to the separable channels. The one-shot distillable dynamic entanglement is bounded by the hypothesis-testing relative entropy of dynamic entanglement. When it comes to the catalytic scenario where additional dynamic entanglement resource is supplied and returned back after the free superchannel, we find that the one-shot catalytic dynamic entanglement cost is bounded by the generalized log-robustness of channels with respect to the set of separable channels. Finally, in the appendices, we investigate the asymptotic scenario, using the liberal dynamic entanglement cost of a bipartite quantum channel, which features the liberal smoothing instead of the diamond norm smoothing. It is shown that the quantity is equal to the liberal regularized relative entropy of channels minimized over the separable channels.

Appendix

A. Liberal Dynamic Entanglement Cost of a Bipartite Channel

There are several alternative ways of smoothing in channel resource theories that could be utilized in the study of the asymptotic equipartition properties [45]. For a quantum channel $\mathcal{N}_A$ and a quantum state $\varphi_{A_0R_0}$, we denote the ε-diamond ball and the ε-liberal ball as

$$B_{\varepsilon}(\mathcal{N}_A) \coloneqq \left\{\mathcal{N}'_A \in \text{CPTP}(A) : \frac{1}{2}\left\|\mathcal{N}'_A - \mathcal{N}_A\right\|_{\diamond} \leq \varepsilon\right\}, \tag{82}$$

$$\begin{aligned} B^{\varphi}_{\varepsilon}(\mathcal{N}_A) \coloneqq \Bigg\{&\mathcal{N}'_A \in \text{CPTP}(A) : \\ &\frac{1}{2}\left\|\mathcal{N}'_A(\varphi_{A_0R_0}) - \mathcal{N}_A(\varphi_{A_0R_0})\right\|_1 \leq \varepsilon\Bigg\}. \end{aligned} \tag{83}$$

Observe that $B_{\varepsilon}(\mathcal{N}_A) \subset \cap_{\varphi_{A_0R_0}} B^{\varphi}_{\varepsilon}(\mathcal{N}_A)$. For a set of free resource $\mathcal{F}$, the relevant liberal quantities are defined as follows:

$$LR^{\varepsilon,\varphi}_{\mathcal{F}}(\mathcal{N}_A) \coloneqq \min_{\mathcal{N}'_A\in B^{\varphi}_{\varepsilon}(\mathcal{N}_A)} LR_{\mathcal{F}}(\mathcal{N}'_A), \tag{84}$$

$$LR^{\varepsilon}_{\mathcal{F}}(\mathcal{N}_A) \coloneqq \max_{\varphi_{A_0R_0}} \min_{\mathcal{N}'_A\in B^{\varphi}_{\varepsilon}(\mathcal{N}_A)} LR_{\mathcal{F}}(\mathcal{N}'_A), \tag{85}$$

$$LR^{\varepsilon,n}_{\mathcal{F}}(\mathcal{N}_A) \coloneqq \frac{1}{n}\max_{\varphi_{A_0R_0}} LR^{\varepsilon,\varphi^{\otimes n}}_{\mathcal{F}}(\mathcal{N}_A^{\otimes n}), \tag{86}$$

$$LR^{(\infty)}_{\mathcal{F}}(\mathcal{N}_A) \coloneqq \lim_{\varepsilon\to 0^+}\liminf_{n\to\infty} LR^{\varepsilon,n}_{\mathcal{F}}(\mathcal{N}_A). \tag{87}$$

The liberal regularized relative entropy of a channel $\mathcal{N}_A$ with respect to a free resource $\mathcal{F}$ is defined as

$$\begin{aligned} D^{(\infty)}_{\mathcal{F}}(\mathcal{N}_A) \coloneqq \lim_{n\to\infty}\frac{1}{n}\max_{\varphi_{A_0R_0}}\min_{\mathcal{M}\in\mathcal{F}} \\ D\left(\mathcal{N}_A^{\otimes n}\left(\varphi^{\otimes n}_{A_0R_0}\right)\middle\|\mathcal{M}_A^{\otimes n}\left(\varphi^{\otimes n}_{A_0R_0}\right)\right). \end{aligned} \tag{88}$$

It is shown in [45] that the asymptotic equipartition property holds as

$$LR_{\mathcal{F}}^{(\infty)}(\mathcal{N}_A) = D_{\mathcal{F}}^{(\infty)}(\mathcal{N}_A). \tag{89}$$

Definition A.1: Given $\varepsilon \geq 0$, the ε-liberal one-shot dynamic entanglement cost of a bipartite quantum channel $\mathcal{N}_{AB}$ under SEPPSC is defined as

$$E_{C_l,\text{SEPPSC}}^{(1),\varepsilon}(\mathcal{N}_{AB}) \coloneqq \max_{\varphi_{A_0A_0'B_0B_0'}} E_{C_l,\text{SEPPSC}}^{(1),\varepsilon,\varphi}(\mathcal{N}_{AB}), \tag{90}$$

where

$$E_{C_l,\text{SEPPSC}}^{(1),\varepsilon,\varphi}(\mathcal{N}_{AB}) \coloneqq \min_{\mathcal{N}'_{AB}\in B_\varepsilon^\varphi(\mathcal{N}_{AB})} E_{C,\text{SEPPSC}}^{(1),0}(\mathcal{N}'_{AB}). \tag{91}$$

The liberal (asymptotic) dynamic entanglement cost of a bipartite quantum channel $\mathcal{N}_{AB}$ under SEPPSC is defined as

$$E_{C_l,\text{SEPPSC}}(\mathcal{N}_{AB}) \coloneqq \lim_{\varepsilon\to 0^+} \liminf_{n\to\infty} \frac{1}{n} \max_{\varphi_{A_0A_0'B_0B_0'}} E_{C_l,\text{SEPPSC}}^{(1),\varepsilon,\varphi^{\otimes n}}(\mathcal{N}_{AB}^{\otimes n}). \tag{92}$$

While an operational meaning of the above quantity is missing yet, it is given by the liberal regularized relative entropy:

Theorem A.1: The liberal dynamic entanglement cost of a bipartite quantum channel $\mathcal{N}_{AB}$ is given by

$$E_{C_l,\text{SEPPSC}}(\mathcal{N}_{AB}) = D_{\text{SepC}}^{(\infty)}(\mathcal{N}_{AB}). \tag{93}$$

Proof: From Theorem III.1, for any $\varepsilon \geq 0$ and φ it holds that

$$LR_{\text{SepC}}^{\varepsilon,\varphi}(\mathcal{N}_{AB}) \leq E_{C_l,\text{SEPPSC}}^{(1),\varepsilon,\varphi}(\mathcal{N}_{AB}) \leq LR_{\text{SepC}}^{\varepsilon,\varphi}(\mathcal{N}_{AB}) + 2. \tag{94}$$

The asymptotic equipartition property leads to the conclusion. □

B. A Few Technical Results

Proposition A.2: Let $|\Phi^K\rangle_{A_0B_0} = \frac{1}{\sqrt{K}}\sum_{i=0}^{K-1}|ii\rangle_{A_0B_0}$ be a K-maximally entangled state where $|A_0| \equiv \dim A_0 \geq K$ and $|B_0| \equiv \dim B_0 \geq K$. We have that

$$\max_{\sigma\in\text{SepD}(A_0:B_0)} \operatorname{Tr}\Phi_{A_0B_0}^K\sigma_{A_0B_0} = \frac{1}{K}. \tag{95}$$

Proof: A separable state $\sigma_{A_0B_0}$ can be written as a convex sum of pure product states $\sigma_{A_0B_0} = \sum_\alpha p_\alpha\phi_\alpha\otimes\psi_\alpha$:

$$\begin{aligned}
&\operatorname{Tr}\Phi_{A_0B_0}^K\sigma_{A_0B_0}\\
&= \frac{1}{K}\sum_{i,j=0}^{K-1}\sum_\alpha p_\alpha\langle i|\phi_\alpha\rangle\langle i|\psi_\alpha\rangle\langle\phi_\alpha|j\rangle\langle\psi_\alpha|j\rangle\\
&= \frac{1}{K}\sum_\alpha p_\alpha\left|\sum_{i=0}^{K-1}\langle i|\phi_\alpha\rangle\langle i|\psi_\alpha\rangle\right|^2\\
&\leq \frac{1}{K}\sum_\alpha p_\alpha\left\{\sum_{i=0}^{K-1}|\langle i|\phi_\alpha\rangle|^2\right\}\left\{\sum_{i=0}^{K-1}|\langle i|\psi_\alpha\rangle|^2\right\}\\
&\leq \frac{1}{K},
\end{aligned} \tag{96}$$

where the Cauchy-Schwarz inequality is used for the first inequality. □

Theorem A.3 [68]: Let $\mathcal{N}_A$ and $\mathcal{M}_A$ be quantum channels, and $J^{\mathcal{N}_A}$ and $J^{\mathcal{M}_A}$ be their (normalized) Choi matrices, respectively. It holds that

$$\frac{1}{|A|}\left\|\mathcal{N}_A-\mathcal{M}_A\right\|_\diamond \leq \left\|J^{\mathcal{N}_A}-J^{\mathcal{M}_A}\right\|_1 \leq \left\|\mathcal{N}_A-\mathcal{M}_A\right\|_\diamond. \tag{97}$$

Proof: The second inequality follows from the definition of the diamond norm. For the first inequality, let $\Psi_{A_0R_0}$ be the optimal pure state for the diamond distance as

$$\left\|\mathcal{N}_A-\mathcal{M}_A\right\|_\diamond = \left\|\mathcal{N}_A(\Psi_{A_0R_0})-\mathcal{M}_A(\Psi_{A_0R_0})\right\|_1, \tag{98}$$

where $|A_0| = |R_0|$. One can denote $|\Psi\rangle_{A_0R_0} = I_{A_0}\otimes X_{R_0}|\phi^+\rangle_{A_0R_0}$, where $|\phi^+\rangle_{A_0R_0} = \sum_{i=0}^{|A_0|-1}|ii\rangle_{A_0R_0}$ and the operator X_{R_0} satisfies $\operatorname{Tr}_{A_0R_0}\Psi_{A_0R_0} = 1 = \operatorname{Tr}_{R_0}X_{R_0}^\dagger X_{R_0} = \|X_{R_0}\|_2^2$. With $\Phi_{A_0R_0}^+ = \frac{\phi_{A_0R_0}^+}{|A_0|}$, we have

$$\begin{aligned}
&\left\|\mathcal{N}_A-\mathcal{M}_A\right\|_\diamond\\
&= \left\|\mathcal{N}_A(\Psi_{A_0R_0})-\mathcal{M}_A(\Psi_{A_0R_0})\right\|_1\\
&= \left\|(\mathcal{N}_A-\mathcal{M}_A)\left(I_{A_0}\otimes X_{R_0}\cdot|A_0|\,\Phi_{A_0R_0}^+\cdot I_{A_0}\otimes X_{R_0}^\dagger\right)\right\|_1\\
&= |A_0|\left\|I_{A_0}\otimes X_{R_0}\cdot\left(J^{\mathcal{N}_A}-J^{\mathcal{M}_A}\right)\cdot I_{A_0}\otimes X_{R_0}^\dagger\right\|_1\\
&\leq |A_0|\left\|X_{R_0}\right\|_\infty\left\|X_{R_0}^\dagger\right\|_\infty\left\|J^{\mathcal{N}_A}-J^{\mathcal{M}_A}\right\|_1\\
&\leq |A_0|\left\|X_{R_0}\right\|_2\left\|X_{R_0}^\dagger\right\|_2\left\|J^{\mathcal{N}_A}-J^{\mathcal{M}_A}\right\|_1\\
&\leq |A_0|\left\|J^{\mathcal{N}_A}-J^{\mathcal{M}_A}\right\|_1,
\end{aligned} \tag{99}$$

where we used the Hölder inequality for the first inequality. □

Theorem A.4: Let $\mathcal{N}_A$ and $\mathcal{M}_A$ be quantum channels. Given $\varepsilon \geq 0$, we have that

$$\begin{aligned}
&\frac{1}{2}\left\|\mathcal{N}_A-\mathcal{M}_A\right\|_\diamond \leq \varepsilon \implies\\
&\min_{\Psi_{A_0R_0}} F(\mathcal{N}_A(\Psi_{A_0R_0}),\mathcal{M}_A(\Psi_{A_0R_0})) \geq (1-\varepsilon)^2 \geq 1-2\varepsilon.
\end{aligned} \tag{100}$$

Conversely, it follows that

$$\begin{aligned}
&\min_{\Psi_{A_0R_0}} F(\mathcal{N}_A(\Psi_{A_0R_0}),\mathcal{M}_A(\Psi_{A_0R_0})) \geq 1-\varepsilon\\
&\qquad\qquad \implies \frac{1}{2}\left\|\mathcal{N}_A-\mathcal{M}_A\right\|_\diamond \leq \sqrt{\varepsilon}.
\end{aligned} \tag{101}$$

Proof: Both follow from the Fuchs-van de Graaf inequality, while $(1-\varepsilon)^2 = 1-2\varepsilon+\varepsilon^2 \geq 1-2\varepsilon$ for $\varepsilon \geq 0$. □

Proposition A.5: Let $\mathcal{N}_A$ and $\mathcal{M}_A$ be quantum channels, and $J^{\mathcal{N}_A}$ and $J^{\mathcal{M}_A}$ be their (normalized) Choi matrices, respectively. If $F\left(J^{\mathcal{N}_A},J^{\mathcal{M}_A}\right) \geq 1-\varepsilon$, then $\frac{1}{2}\left\|\mathcal{N}_A-\mathcal{M}_A\right\|_\diamond \leq n\sqrt{\varepsilon}$.

Proof: From Theorem A.3 and the Fuchs-van der Graaf inequality, it follows that

$$\begin{aligned}\frac{1}{2}\left\|\mathcal{N}_A-\mathcal{M}_A\right\|_\diamond &\leq n\frac{1}{2}\left\|J^{\mathcal{N}_A}-J^{\mathcal{M}_A}\right\|_1\\ &\leq n\sqrt{1-F\left(J^{\mathcal{N}_A},J^{\mathcal{M}_A}\right)}\\ &\leq n\sqrt{\varepsilon}.\end{aligned}$$

□

NOTE ADDED

During the completion of this manuscript, we became aware of two independent works on dynamic resource theories: Regula and Takagi [69] formulated one-shot manipulation of dynamic resources in a general setting, while Yuan, *et al.* [70] also investigated one-shot distillation and dilution of dynamic resources in a general setting.

REFERENCES

[1] E. Chitambar and G. Gour, "Quantum resource theories," *Rev. Mod. Phys.*, vol. 91, no. 2, Apr. 2019, Art. no. 025001.

[2] T. Baumgratz, M. Cramer, and M. B. Plenio, "Quantifying coherence," *Phys. Rev. Lett.*, vol. 113, no. 14, Sep. 2014, Art. no. 140401. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.113.140401

[3] A. Winter and D. Yang, "Operational resource theory of coherence," *Phys. Rev. Lett.*, vol. 116, no. 12, Mar. 2016, Art. no. 120404. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.116.120404

[4] A. Streltsov, G. Adesso, and M. B. Plenio, "Colloquium: Quantum coherence as a resource," *Rev. Mod. Phys.*, vol. 89, no. 4, Oct. 2017, Art. no. 041003. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.89.041003

[5] M. B. Plenio and S. Virmani, "An introduction to entanglement measures," *Quantum Inf. Comput.*, vol. 7, nos. 1–2, pp. 1–51, Jan. 2007.

[6] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, "Quantum entanglement," *Rev. Mod. Phys.*, vol. 81, p. 865, Jun. 2009.

[7] J. Åberg, "Quantifying superposition," Dec. 2006, *arXiv:quant-ph/0612146*. [Online]. Available: http://arxiv.org/abs/quant-ph/0612146

[8] T. Theurer, N. Killoran, D. Egloff, and M. B. Plenio, "Resource theory of superposition," *Phys. Rev. Lett.*, vol. 119, no. 23, Dec. 2017, Art. no. 230401. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.119.230401

[9] G. Gour and R. W. Spekkens, "The resource theory of quantum reference frames: Manipulations and monotones," *New J. Phys.*, vol. 10, no. 3, Mar. 2008, Art. no. 033023.

[10] L. Lami, B. Regula, X. Wang, R. Nichols, A. Winter, and G. Adesso, "Gaussian quantum resource theories," *Phys. Rev. A, Gen. Phys.*, vol. 98, no. 2, Aug. 2018, Art. no. 022335.

[11] K. C. Tan, T. Volkoff, H. Kwon, and H. Jeong, "Quantifying the coherence between coherent states," *Phys. Rev. Lett.*, vol. 119, no. 19, Nov. 2017, Art. no. 190405. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.119.190405

[12] B. Yadin, F. C. Binder, J. Thompson, V. Narasimhachar, M. Gu, and M. S. Kim, "Operational resource theory of continuous-variable nonclassicality," *Phys. Rev. X*, vol. 8, no. 4, Dec. 2018, Art. no. 041038.

[13] G. Ferrari, L. Lami, T. Theurer, and M. B. Plenio, "Asymptotic state transformations of continuous variable resources," 2020, *arXiv:2010.00044*. [Online]. Available: http://arxiv.org/abs/2010.00044

[14] N. Killoran, M. Cramer, and M. B. Plenio, "Extracting entanglement from identical particles," *Phys. Rev. Lett.*, vol. 112, no. 15, Apr. 2014, Art. no. 150501.

[15] N. Killoran, F. E. S. Steinhoff, and M. B. Plenio, "Converting nonclassicality into entanglement," *Phys. Rev. Lett.*, vol. 116, no. 8, Feb. 2016, Art. no. 080402.

[16] B. Morris, B. Yadin, M. Fadel, T. Zibold, P. Treutlein, and G. Adesso, "Entanglement between identical particles is a useful and consistent resource," *Phys. Rev. X*, vol. 10, no. 4, Oct. 2020, Art. no. 041012.

[17] F. G. S. L. Brandão, M. Horodecki, J. Oppenheim, J. M. Renes, and R. W. Spekkens, "Resource theory of quantum states out of thermal equilibrium," *Phys. Rev. Lett.*, vol. 111, no. 25, Dec. 2013, Art. no. 250404. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.111.250404

[18] M. Horodecki and J. Oppenheim, "Fundamental limitations for quantum and nanoscale thermodynamics," *Nature Commun.*, vol. 4, no. 1, p. 2059, Jun. 2013. [Online]. Available: http://www.nature.com/ncomms/2013/130626/ncomms3059/full/ncomms3059.html#supplementary-information

[19] F. G. S. L. Brandão, M. Horodecki, N. Ng, J. Oppenheim, and S. Wehner, "The second laws of quantum thermodynamics," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 11, pp. 3275–3279, Mar. 2015. [Online]. Available: http://www.pnas.org/content/112/11/3275

[20] A. Peres, "Separability criterion for density matrices," *Phys. Rev. Lett.*, vol. 77, no. 8, pp. 1413–1415, Aug. 1996. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.77.1413

[21] M. Horodecki, P. Horodecki, and R. Horodecki, "Separability of mixed states: Necessary and sufficient conditions," *Phys. Lett. A*, vol. 223, nos. 1–2, pp. 1–8, Nov. 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0375960196007062

[22] C. H. Bennett, H. J. Bernstein, S. Popescu, and B. Schumacher, "Concentrating partial entanglement by local operations," *Phys. Rev. A, Gen. Phys.*, vol. 53, no. 4, p. 2046, 1996. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevA.53.2046

[23] M. A. Nielsen, "Conditions for a class of entanglement transformations," *Phys. Rev. Lett.*, vol. 83, no. 2, pp. 436–439, Jul. 1999. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.83.436

[24] D. Jonathan and M. B. Plenio, "Minimal conditions for local pure-state entanglement manipulation," *Phys. Rev. Lett.*, vol. 83, no. 7, pp. 1455–1458, Aug. 1999. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.83.1455

[25] G. Vidal, D. Jonathan, and M. A. Nielsen, "Approximate transformations and robust manipulation of bipartite pure-state entanglement," *Phys. Rev. A, Gen. Phys.*, vol. 62, no. 1, Jun. 2000, Art. no. 012304. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevA.62.012304

[26] K. M. R. Audenaert and M. B. Plenio, "When are correlations quantum?—Verification and quantification of entanglement by simple measurements," *New J. Phys.*, vol. 8, no. 11, p. 266, 2006.

[27] J. Eisert, F. G. S. L. Brandão, and K. M. R. Audenaert, "Quantitative entanglement witnesses," *New J. Phys.*, vol. 9, no. 3, p. 46, 2007.

[28] O. Gühne, M. Reimpell, and R. F. Werner, "Estimating entanglement measures in experiments," *Phys. Rev. Lett.*, vol. 98, no. 11, Mar. 2007, Art. no. 110502.

[29] B. P. Lanyon *et al.*, "Efficient tomography of a quantum many-body system," *Nature Phys.*, vol. 13, no. 12, pp. 1158–1162, 2017.

[30] J. Eisert, K. Jacobs, P. Papadopoulos, and M. B. Plenio, "Optimal local implementation of nonlocal quantum gates," *Phys. Rev. A, Gen. Phys.*, vol. 62, no. 5, Oct. 2000, Art. no. 052317. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.62.052317

[31] D. Collins, N. Linden, and S. Popescu, "Nonlocal content of quantum operations," *Phys. Rev. A, Gen. Phys.*, vol. 64, no. 3, Aug. 2001, Art. no. 032302.

[32] B. Coecke, T. Fritz, and R. W. Spekkens, "A mathematical theory of resources," *Inf. Comput.*, vol. 250, pp. 59–86, Oct. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0890540116000353

[33] G. Gour and M. M. Wilde, "Entropy of a quantum channel," Aug. 2018, *arXiv:1808.06980*. [Online]. Available: http://arxiv.org/abs/1808.06980

[34] T. Theurer, D. Egloff, L. Zhang, and M. B. Plenio, "Quantifying operations with an application to coherence," *Phys. Rev. Lett.*, vol. 122, no. 19, May 2019, Art. no. 190405.

[35] Z.-W. Liu and A. Winter, "Resource theories of quantum channels and the universal role of resource erasure," Apr. 2019, *arXiv:1904.04201*. [Online]. Available: http://arxiv.org/abs/1904.04201

[36] Y. Liu and X. Yuan, "Operational resource theory of quantum channels," *Phys. Rev. Res.*, vol. 2, no. 1, Feb. 2020, Art. no. 012035.

[37] G. Saxena, E. Chitambar, and G. Gour, "Dynamical resource theory of quantum coherence," *Phys. Rev. Res.*, vol. 2, no. 2, Jun. 2020, Art. no. 023298. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevResearch.2.023298

[38] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, "Fundamental limits of repeaterless quantum communications," *Nature Commun.*, vol. 8, no. 1, p. 15043, Apr. 2017. [Online]. Available: https://www.nature.com/articles/ncomms15043

[39] T. Theurer, S. Satyajit, and M. B. Plenio, "Quantifying dynamical coherence with dynamical entanglement," *Phys. Rev. Lett.*, vol. 125, no. 13, Sep. 2020, Art. no. 130401.

[40] G. Gour and C. M. Scandolo, "The entanglement of a bipartite channel," Jul. 2019, *arXiv:1907.02552*. [Online]. Available: http://arxiv.org/abs/1907.02552

[41] G. Gour and C. M. Scandolo, "Dynamical entanglement," *Phys. Rev. Lett.*, vol. 125, no. 18, Oct. 2020, Art. no. 180505. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.125.180505

[42] X. Wang and M. M. Wilde, "Cost of quantum entanglement simplified," *Phys. Rev. Lett.*, vol. 125, no. 4, Jul. 2020, Art. no. 040502. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.125.040502

[43] X. Wang and M. M. Wilde, "Exact entanglement cost of quantum states and channels under PPT-preserving operations," Sep. 2018, *arXiv:1809.09592*. [Online]. Available: http://arxiv.org/abs/1809.09592

[44] S. Bäuml, S. Das, X. Wang, and M. M. Wilde, "Resource theory of entanglement for bipartite quantum channels," Jul. 2019, *arXiv:1907.04181*. [Online]. Available: http://arxiv.org/abs/1907.04181

[45] G. Gour and A. Winter, "How to quantify a dynamical quantum resource," *Phys. Rev. Lett.*, vol. 123, no. 15, Oct. 2019, Art. no. 150401. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.123.150401

[46] A. Y. Kitaev, "Quantum computations: Algorithms and error correction," *Russian Math. Surv.*, vol. 52, no. 6, pp. 1191–1249, Dec. 1997. [Online]. Available: http://stacks.iop.org/0036-0279/52/i=6/a=R02?key=crossref.9a3dd3d906e4a9338822c6992ae26e8a

[47] V. Paulsen, *Completely Bounded Maps and Operator Algebras*. Cambridge, U.K.: Cambridge Univ. Press, 2003. [Online]. Available: http://ebooks.cambridge.org/ref/id/CBO9780511546631

[48] F. G. S. L. Brandão and M. B. Plenio, "A reversible theory of entanglement and its relation to the second law," *Commun. Math. Phys.*, vol. 295, no. 3, pp. 829–851, May 2010.

[49] A. W. Harrow and M. A. Nielsen, "Robustness of quantum gates in the presence of noise," *Phys. Rev. A, Gen. Phys.*, vol. 68, no. 1, Jul. 2003, Art. no. 012308. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.68.012308

[50] Z.-W. Liu, K. Bu, and R. Takagi, "One-shot operational quantum resource theory," *Phys. Rev. Lett.*, vol. 123, no. 2, Jul. 2019, Art. no. 020401. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.123.020401

[51] G. Chiribella, G. M. D'Ariano, and P. Perinotti, "Transforming quantum operations: Quantum supermaps," *Europhys. Lett.*, vol. 83, no. 3, p. 30004, Aug. 2008.

[52] G. Gour, "Comparison of quantum channels by superchannels," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5880–5904, Sep. 2019.

[53] G. Vidal and R. Tarrach, "Robustness of entanglement," *Phys. Rev. A, Gen. Phys.*, vol. 59, no. 1, pp. 141–155, Jan. 1999. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.59.141

[54] M. Steiner, "Generalized robustness of entanglement," *Phys. Rev. A, Gen. Phys.*, vol. 67, no. 5, May 2003, Art. no. 054305. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.67.054305

[55] B. Regula, "Convex geometry of quantum resource quantification," *J. Phys. A, Math. Theor.*, vol. 51, no. 4, Dec. 2017, Art. no. 045303.

[56] R. Takagi and B. Regula, "General resource theories in quantum mechanics and beyond: Operational characterization via discrimination tasks," *Phys. Rev. X*, vol. 9, no. 3, Sep. 2019, Art. no. 031053. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.9.031053

[57] L. Lami, B. Regula, R. Takagi, and G. Ferrari, "Framework for resource quantification in infinite-dimensional general probabilistic theories," *Phys. Rev. A, Gen. Phys.*, vol. 103, no. 3, Mar. 2021, Art. no. 032424. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.103.032424

[58] B. Regula, L. Lami, G. Ferrari, and R. Takagi, "Operational quantification of continuous-variable quantum resources," *Phys. Rev. Lett.*, vol. 126, no. 11, Mar. 2021, Art. no. 110403. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.126.110403

[59] M. A. Nielsen *et al.*, "Quantum dynamics as a physical resource," *Phys. Rev. A, Gen. Phys.*, vol. 67, no. 5, May 2003, Art. no. 052301.

[60] M. M. Wolf. (2012). *Quantum Channels and Operations—Guided Tour*. [Online]. Available: http://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MichaelWolf/QChannelLecture.pdf

[61] M. Horodecki and P. Horodecki, "Reduction criterion of separability and limits for a class of distillation protocols," *Phys. Rev. A, Gen. Phys.*, vol. 59, no. 6, pp. 4206–4216, Jun. 1999. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevA.59.4206

[62] F. G. S. L. Brandão and N. Datta, "One-shot rates for entanglement manipulation under non-entangling maps," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1754–1760, Mar. 2011.

[63] T. Cooney, M. Mosonyi, and M. M. Wilde, "Strong converse exponents for a quantum channel discrimination problem and quantum-feedback-assisted communication," *Commun. Math. Phys.*, vol. 344, no. 3, pp. 797–829, Jun. 2016.

[64] X. Yuan, "Hypothesis testing and entropies of quantum channels," *Phys. Rev. A, Gen. Phys.*, vol. 99, no. 3, Mar. 2019, Art. no. 032317.

[65] C. A. Fuchs and J. van de Graaf, "Cryptographic distinguishability measures for quantum-mechanical states," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1216–1227, May 1999.

[66] J. Watrous, *The Theory of Quantum Information*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, Apr. 2018.

[67] F. G. S. L. Brandão and M. B. Plenio, "Entanglement theory and the second law of thermodynamics," *Nature Phys.*, vol. 4, no. 11, pp. 873–877, Nov. 2008.

[68] J. Watrous. (2011). *Is There Any Connection Between the Diamond Norm and the Distance of the Associated States?* [Online]. Available: https://cstheory.stackexchange.com/q/4920

[69] B. Regula and R. Takagi, "One-shot manipulation of dynamical quantum resources," Dec. 2020, *arXiv:2012.02215*. [Online]. Available: http://arxiv.org/abs/2012.02215

[70] X. Yuan, P. Zeng, M. Gao, and Q. Zhao, "One-shot dynamical resource theory," Dec. 2020, *arXiv:2012.02781*. [Online]. Available: http://arxiv.org/abs/2012.02781

Ho-Joon Kim received the Doctor of Philosophy degree in physics from the Korea Advanced Institute of Science and Technology (KAIST) in 2010. He is currently a Research Professor with Kyung Hee University. He is interested in quantum information theory, quantum computation, and quantum optics.

Soojoon Lee received the Ph.D. degree from the Department of Mathematical Sciences, Seoul National University in 2002, with a focus on quantum computational algorithms. After postdoctoral positions at the Statistical Research Center for Complex Systems, Seoul National University and School of Computational Sciences, Korea Institute for Advanced Study (KIAS), he joined the Department of Mathematics, Kyung Hee University, in 2004. He is currently a Professor of mathematics with Kyung Hee University and an Associate Member with KIAS. He is interested in quantum algorithms and quantum information theory, including quantum communication and entanglement theory.

Ludovico Lami received the M.Sc. degree in physics from the Università di Pisa, Pisa, Italy, in 2014, the Diploma degree in physics from the Scuola Normale Superiore, Pisa, in 2015, and the Ph.D. degree from the Department de Física, Universitat Autònoma de Barcelona, Barcelona, Spain, in 2017. He is currently an Alexander von Humboldt Research Fellow with the University of Ulm. His research interests lie in quantum information, continuous variable, and foundational aspects of quantum physics.

Martin B. Plenio received the Dr. rer. nat. degree in physics from the Georg-August-Universität Gottingen in 1994. He is currently an Alexander von Humboldt-Professor and the Director of the Institute of Theoretical Physics, Universiät Ulm. He works in the fields of quantum information theory, quantum technologies, and quantum biology.

Path identity as a source of high-dimensional entanglement

Jaroslav Kysela[a,b,1,2], Manuel Erhard[a,b,1,2], Armin Hochrainer[a,b], Mario Krenn[a,b,c], and Anton Zeilinger[a,b,2]

[a]Faculty of Physics, Vienna Center for Quantum Science & Technology, University of Vienna, 1090 Vienna, Austria; [b]Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, 1090 Vienna, Austria; and [c]Department of Chemistry & Computer Science, University of Toronto, Toronto, ON M5S 3H6, Canada

Contributed by Anton Zeilinger, September 3, 2020 (sent for review June 23, 2020; reviewed by Mohamed Bourennane and Sergey Kulik)

We present an experimental demonstration of a general entanglement-generation framework, where the form of the entangled state is independent of the physical process used to produce the particles. It is the indistinguishability of multiple generation processes and the geometry of the setup that give rise to the entanglement. Such a framework, termed entanglement by path identity, exhibits a high degree of customizability. We employ one class of such geometries to build a modular source of photon pairs that are high-dimensionally entangled in their orbital angular momentum. We demonstrate the creation of three-dimensionally entangled states and show how to incrementally increase the dimensionality of entanglement. The generated states retain their quality even in higher dimensions. In addition, the design of our source allows for its generalization to various degrees of freedom and even for the implementation in integrated compact devices. The concept of entanglement by path identity itself is a general scheme and allows for construction of sources producing also customized states of multiple photons. We therefore expect that future quantum technologies and fundamental tests of nature in higher dimensions will benefit from this approach.

entanglement by path identity | high-dimensional entanglement | path indistinguishability | orbital angular momentum

The transition from two- to multidimensional entangled quantum systems brings about radical improvements in the distribution and processing of quantum information. Such systems play an important role in secure high-dimensional superdense coding schemes (1–3); they offer improved noise resistance and increased security against eavesdropping (4, 5); and they are beneficial or even indispensable for fundamental experiments, such as tests of local realism (6–9) or the prospect of teleportation of the entire information stored in a photonic system (10–12). Various degrees of freedom, such as frequency (13), time bin (14–16), and path (17, 18), have been employed so far for the generation of high-dimensionally entangled states. In this paper, we present an experimental proof-of-principle demonstration of a conceptually different framework of generating high-dimensionally entangled states. Multiple spontaneous parametric down-conversion (SPDC) processes are employed, but none of them individually produces entanglement. The entanglement is built in a manner, where not intrinsic properties of a photon-production process, but rather the geometry of the setup governs the structure of the final entangled state. This method amounts to the concept known as entanglement by path identity (19, 20), which was discovered recently with the help of a computer program (21). Utilizing this concept leads to a simple yet versatile design of a source of high-dimensional entanglement. In the following, we present the experimental implementation of this source adapted to the orbital angular momentum (OAM) of photons. Nevertheless, the scheme is not linked to a specific degree of freedom and is valid for other degrees of freedom as well.

The OAM of photons is an in principle unbounded discrete quantity and as such has been used extensively (22–26) to prepare high-dimensionally entangled photonic states. In the traditional way, the OAM-entangled photon pairs are produced in a single SPDC process (27). Albeit convenient, this process exhibits several drawbacks. For example, photon pairs generated in this way have a nonuniform distribution of OAM (28–31). The maximally entangled states can then be generated either by postprocessing techniques, such as Procrustean filtering (32, 33), or by preprocessing of the pump beam. In a recently demonstrated approach (34, 35), a superposition of OAM modes is imprinted by holograms into the pump beam, which translates via down-conversion into maximally entangled states of two photons.

Our technique offers several important advantages over the traditional approach. The source of entangled photon pairs enables us to engineer the state for our needs as both phases and magnitudes in a high-dimensional quantum state can be adjusted completely arbitrarily. One is not limited by the conditions of the employed SPDC processes. This way, various families of states can be produced, such as high-dimensional maximally entangled Bell states that are demanded by applications such as high-dimensional quantum dense coding (36), entanglement swapping (37), or quantum teleportation (38). By proper adjustment of the state's magnitudes the nonmaximally entangled states maximizing the violation of high-dimensional Bell inequalities (6, 39) can be also produced. The experimental implementation

Significance

Quantum entanglement amounts to an extremely strong link between two distant particles, a link so strong that it eludes any classical description and so unsettling that Albert Einstein described it as "spooky action at a distance." Today, entanglement is not only a subject of fundamental research, but also a workhorse of emerging quantum technologies. In our current work we experimentally demonstrate a completely different method of entanglement generation. Unlike many traditional methods, where entanglement arises due to conservation of a physical quantity, such as momentum, in our method it is rather a consequence of indistinguishability of several particle-generating processes. This approach, where each process effectively adds one dimension to the entangled state, allows for a high degree of customizability.

Author contributions: M.E., A.H., and M.K. designed research; J.K. and M.E. performed research; J.K. and M.E. analyzed data; J.K., M.E., A.H., M.K., and A.Z. wrote the paper; and A.Z. initiated and supervised research.

Reviewers: M.B., Stockholm University; and S.K., Moscow State University.

Competing interest statement: A.Z. and M.B. are coauthors on a 2018 paper of a large community proposing a space experiment. They did not collaborate directly on this work.

[1]J.K. and M.E. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: anton.zeilinger@univie.ac.at, jaroslav.kysela@univie.ac.at, or manuel.erhard@univie.ac.at.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2011405117/-/DCSupplemental.

First published October 1, 2020.

of our source has a modular structure, where adding a single module leads to increasing the entanglement dimension by one. High brightness of our source is possible as all photons are produced already in the desired modes and no photons have to be discarded by postselection.

This work is organized as follows. After a brief introduction to the concept of entanglement by path identity, we describe the experimental design of our source. Then we demonstrate the scalability and versatility of our method by generating several different states in two and in three dimensions. We verify the quality of the produced entangled quantum states using quantum state tomography.

Entanglement by Path Identity

Consider a simple experimental setup consisting of two nonlinear crystals that are aligned in series and coherently emit photons via SPDC, as shown in Fig. 1 *A* and *B*. The pump power for both crystals is set sufficiently low such that events when either crystal emits multiple photon pairs as well as events when both crystals each simultaneously generate a photon pair can be neglected. The propagation paths of the down-converted photons coming from the two crystals are carefully overlapped. As a result, once the photon pair leaves the setup, no information can be obtained, not even in principle, in which crystal the pair was created (40–42). The down-conversion processes in both crystals are adjusted such that photon pairs may be emitted only into the fundamental mode $|0,0\rangle$ with zero quanta of OAM*. Importantly, no entanglement is generated by either of the two crystals. (In practice, a small contribution of higher-order OAM modes is also present. For the detailed discussion see *SI Appendix*, *Spiral spectrum*.)

Suppose now that two mode shifters are inserted into the setup. These add an extra quantum of OAM to each photon originating in the first crystal and thus act as the only possible source of which-crystal information. As the down-conversion processes in the two crystals are (apart from the OAM) indistinguishable, the resulting state of a detected photon pair is a coherent superposition

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|0,0\rangle + e^{i\varphi}|1,1\rangle). \qquad [1]$$

In Eq. **1**, φ is the phase between the two SPDC processes imparted by a phase shifter and numbers in ket vectors refer to the OAM quanta of respective photons.

The generation of entangled states as described above is a specific example of the concept termed entanglement by path identity. This concept can be readily generalized for production of high-dimensionally entangled states (19). When the number of crystals in the series is increased to d, and the number of phase and mode shifters is accordingly increased to $d-1$, high-dimensionally entangled states of the following form are produced as

$$|\psi\rangle = \sum_{\ell=0}^{d-1} c_\ell |\ell,\ell\rangle, \qquad [2]$$

where d is the state dimension and c_ℓ are complex amplitudes (Fig. 1 *C* and *D*). The magnitudes of c_ℓ can be set by pumping each crystal independently with properly adjusted power. By using different mode shifters for either of the two photons in a down-converted pair, completely arbitrary states can be created. Interestingly, the widely used cross-crystal scheme is the simplest example of the above approach, where two-particle states are entangled in polarization (43–45).

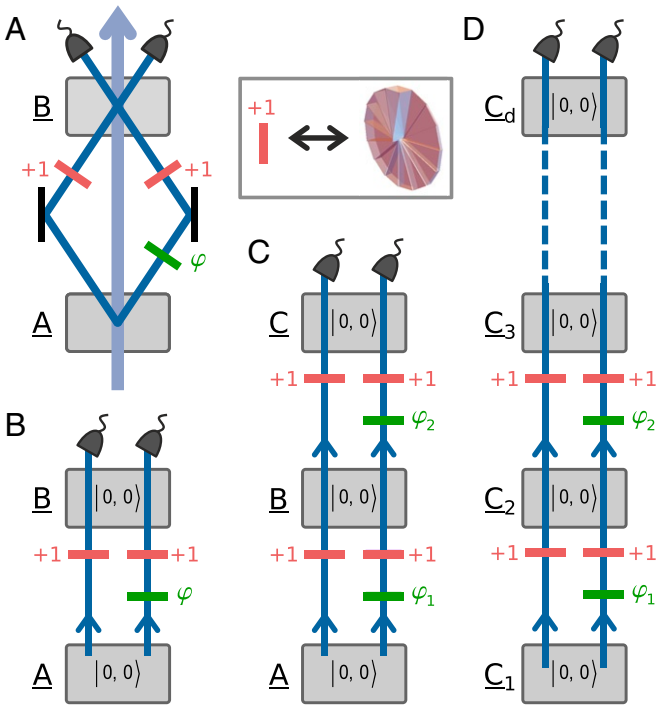


Fig. 1. Basic concept. Gray boxes labeled with underlined uppercase letters represent nonlinear crystals, each pumped coherently and each generating with a small probability a pair of photons via an SPDC process. Each generated photon pair is in OAM state $|0,0\rangle$ with a small contribution of higher-order terms. The two down-converted photons then propagate along their paths in the direction indicated by the arrows and acquire phase shifts φ_i as well as additional quanta of OAM due to phase and mode shifters. (*A*) The pump beam, represented by an arrow, gives rise to an SPDC process in crystals A and B. Photons generated in crystal A are reflected into crystal B such that their paths are overlapped with paths of photons generated in crystal B. As a consequence, the two coherent SPDC processes in crystals A and B are indistinguishable and the generated photon pairs leave the setup in a two-dimensionally entangled Bell state $1/\sqrt{2}\,(|0,0\rangle + \exp(i\varphi)|1,1\rangle)$. The quantum of OAM is imparted to the photon by a spiral phase plate shown in *Inset*. (*B*) A schematic picture of the setup in *A*, where the pump beam is not shown. (*C*) The addition of the third crystal to the setup increases the entanglement dimension by one. The resulting state is thus $1/\sqrt{3}\,(|0,0\rangle + \exp(i\bar{\varphi}_1)|1,1\rangle + \exp(i\bar{\varphi}_2)|2,2\rangle)$, where $\bar{\varphi}_1 = \varphi_2$ and $\bar{\varphi}_2 = \varphi_1 + \varphi_2$. (*D*) One can stack multiple setups from *A* to acquire a series of d crystals that produces a d-dimensionally entangled state $1/\sqrt{d}(|0,0\rangle + \exp(i\bar{\varphi}_1)|1,1\rangle + \cdots + \exp(i\bar{\varphi}_{d-1})|d-1,d-1\rangle)$, where the relative phases $\bar{\varphi}_i = \sum_{j=d-i}^{d-1}\varphi_j$ are adjusted by an appropriate choice of phase shifters φ_j. The magnitudes of the individual modes are modified by varying the power with which the respective crystals are pumped.

Setup

The experimental implementation presented here is based on the scheme in Fig. 1*C* with two main modifications. The pump and down-converted beams for each crystal are separated by two Mach–Zehnder interferometers, such that both wavelengths can be manipulated separately. This way, phases as well as magnitudes of individual modes in the quantum state can be adjusted independently. For technical reasons, the down-converted photon pairs were not emitted in a perfectly collinear manner, but had a slight angular deviation of roughly 1°. This leads to a nonperfect operation of the mode shifter, which functions properly only when both photons propagate through its center. As a countermeasure, we place the mode shifter into the pump beam instead of the down-conversion beam. For details refer to *SI Appendix*, *Detailed setup* and *Coherence conditions*.

*Parameters of the two SPDC processes are chosen such that photon properties such as frequency, polarization, and OAM are identical for both crystals and also higher-order OAM modes are highly suppressed.

The setup, presented in Fig. 2, was designed to produce three-dimensionally entangled states. Each dimension in the generated quantum state corresponds to one of three nonlinear crystals A, B, or C in the setup. In Fig. 2 this correspondence is emphasized by enclosing the crystals with associated elements into boxes labeled 1st dim, 2nd dim, and 3rd dim. The laser beam is split into three paths to pump each crystal separately. The pump beam for crystal A possesses zero quanta of OAM and so do the down-converted photons, which exit the crystal in state $|0,0\rangle$. (Apart from the predominant $|0,0\rangle$ component, also effectively negligible contributions of higher-order OAM terms are present in the photons' state, as detailed in *SI Appendix*, *Spiral spectrum*.) The pump beam for crystal B acquires four quanta of OAM due to a spiral phase plate (SPP), which is inserted into the beam and plays the role of the mode shifter. Consequently, each down-converted photon generated in crystal B carries two quanta of OAM and the pair is produced in state $|2,2\rangle$. Similarly, the pump beam for crystal C also acquires four quanta of OAM, but an additional mirror is used to invert the sign of the OAM value from 4 to -4, effectively subtracting eight quanta of OAM. Down-converted photons coming from crystal C are then produced in state $|-2,-2\rangle$. The resulting quantum state reads

$$|\psi\rangle = \alpha\underbrace{|0,0\rangle}_{\text{crystal A}} + \beta e^{i\varphi_1}\underbrace{|2,2\rangle}_{\text{crystal B}} + \gamma e^{i\varphi_2}\underbrace{|-2,-2\rangle}_{\text{crystal C}}. \qquad [3]$$

Magnitudes α, β, and γ of the entangled state can be changed by adjusting the relative pump power for each crystal. The relative phases φ_1 and φ_2 are set by positioning two trombone systems that act as phase shifters. By employing only the first two stages of the setup, namely parts in boxes labeled 1st dim and 2nd dim, two-dimensionally entangled states are created. In ref. 46 a similar experimental setup was used to generate three-dimensional (3D) nonentangled states of photons in Fock representation.

We use type II SPDC in all three crystals. To measure the entangled state, we first deterministically separate the two down-converted photons by a polarizing beam splitter. Two spatial light modulators in combination with single mode fibers are used to perform any projective measurement for OAM modes (27). The single photons are then detected by avalanche photon detectors and simultaneous two-photon events are identified by a coincidence logic.

Finally, the resulting quantum states are characterized by complete quantum state tomography. We use a maximum-likelihood reconstruction technique (47) to estimate the physical density matrices of the detected photon pairs. Also, using the fidelity bound derived in ref. 48, the minimum generated entanglement dimensionality is found.

Experimental Results

The high flexibility of our setup in producing various states is demonstrated in Table 1, where fidelities for different three-dimensionally (and also two-dimensionally) entangled states are presented. These data demonstrate our ability to control the relative phases and magnitudes of the generated quantum states. Most notably, we are able to create three mutually orthogonal and maximally entangled states in three dimensions $|\psi_1\rangle$, $|\psi_2\rangle$, and $|\psi_3\rangle$ with an average fidelity of $87.5 \pm 2.2\%$. These states represent three of nine two-party 3D Bell states, which are important for example in high-dimensional quantum teleportation (38) or high-dimensional superdense coding schemes (1). The orthogonality of these states does not follow directly from the orthogonality of OAM modes, but indeed from differently adjusted phases in the quantum states. Fidelity bounds derived in refs. 48 and 49, which are calculated as a sum of squares of all but the smallest Schmidt coefficients of a given reference state $|\psi_i\rangle$, are used to determine the entanglement dimensionality of the corresponding measured states. When the fidelity F of the experimentally measured density matrix exceeds the associated fidelity bound, the created state is at least three-dimensionally entangled. The presented states $|\psi_1\rangle$ through $|\psi_4\rangle$ are indeed entangled in three dimensions, as their fidelities F satisfy $F > 2/3 \approx 0.67$ and the same is true for $|\psi_5\rangle$ for which $F > 9/11 \approx 0.82$. Likewise, fidelities for two-dimensional (2D) states $|\Phi^+\rangle$ and $|\Phi^-\rangle$ satisfy $F > 1/2$. With the state $|\psi_5\rangle$ we demonstrate the ability to adjust relative magnitudes of terms in the quantum superposition. A nonmaximally entangled state with uneven magnitudes, very similar to $|\psi_5\rangle$, provides the

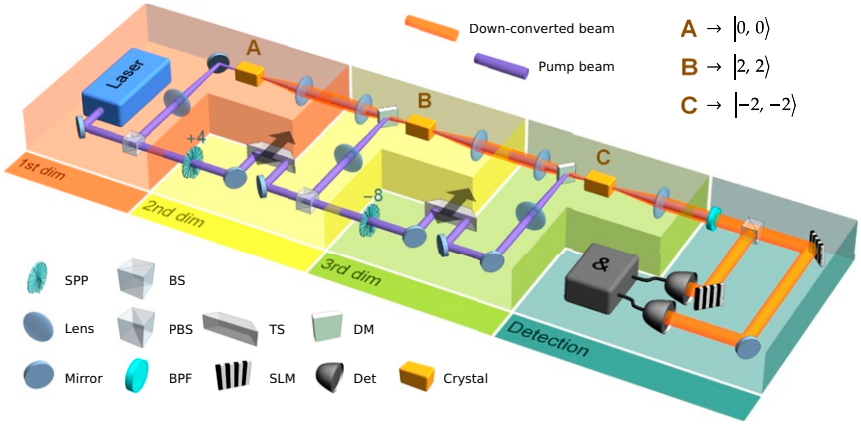


Fig. 2. Experimental setup. Three-dimensional states are created by elements in boxes labeled 1st dim, 2nd dim, and 3rd dim. Three periodically poled KTP crystals A, B, and C are pumped with a continuous-wave laser beam at the central wavelength of 405 nm. Frequency-degenerate down-converted photons created by type II collinear SPDC propagate along identical paths into the detection system shown in box Detection. Photons originating in crystal B are created in $|2,2\rangle$ OAM mode because of a spiral phase plate (SPP $+4$) inserted after the first beam splitter (BS). In addition, photons originating in crystal C are created in $|-2,-2\rangle$ mode due to an extra mirror that effectively works as a -8 mode shifter as is explained in the main text. The pump beam is separated from the down-conversion beam by dichroic mirrors (DM) and a band-pass filter (BPF). Before detection, the two down-converted photons are separated on a polarizing beam-splitter (PBS). The state tomography in the OAM degree of freedom is done by projective measurements (27) where specific holograms are projected on two spatial light modulators (SLMs). The reflected photons are subsequently coupled into single-mode fibers and detected by single-photon detectors (Det). The resulting signals are postprocessed by a coincidence counting module (&). The relative phases φ_1 and φ_2 can be adjusted by phase shifters implemented with trombone systems (TS). The magnitudes of individual terms in the quantum state are controlled by setting the splitting ratio of the beam splitters. For the detailed diagram of the experimental setup see *SI Appendix*.

Table 1. Fidelities $F(|\psi\rangle, \rho) = \mathrm{Tr}(|\psi\rangle\langle\psi|\rho)$ between several two- and three-dimensionally entangled states $|\psi\rangle$ and their experimental realizations ρ

State	Fidelity F
$\lvert\Phi^+\rangle = 1/\sqrt{2}(\lvert 0,0\rangle + \lvert 2,2\rangle)$	0.904 ± 0.005
$\lvert\Phi^-\rangle = 1/\sqrt{2}(\lvert 0,0\rangle - \lvert 2,2\rangle)$	0.891 ± 0.005
$\lvert\psi_1\rangle = \frac{1}{\sqrt{3}}(\lvert 0,0\rangle + \lvert 2,2\rangle + \lvert -2,-2\rangle)$	0.870 ± 0.005
$\lvert\psi_2\rangle = \frac{1}{\sqrt{3}}(\lvert 0,0\rangle + \omega\lvert 2,2\rangle + \omega^{-1}\lvert -2,-2\rangle)$	0.852 ± 0.007
$\lvert\psi_3\rangle = \frac{1}{\sqrt{3}}(\lvert 0,0\rangle + \omega^{-1}\lvert 2,2\rangle + \omega\lvert -2,-2\rangle)$	0.903 ± 0.006
$\lvert\psi_4\rangle = \frac{1}{\sqrt{3}}(\lvert 0,0\rangle - \lvert 2,2\rangle - \lvert -2,-2\rangle)$	0.890 ± 0.004
$\lvert\psi_5\rangle = \frac{1}{\sqrt{22}}(2\lvert 0,0\rangle + 3\lvert 2,2\rangle + 3\lvert -2,-2\rangle)$	0.848 ± 0.008

States $|\psi_1\rangle$, $|\psi_2\rangle$, and $|\psi_3\rangle$ form an orthonormal set of maximally entangled states in three dimensions ($\omega = e^{2\pi i/3}$). State $|\psi_5\rangle$ is a manifestation of our ability to control not only relative phases in the quantum state, but also relative magnitudes. The error estimates are calculated by propagation of Poissonian statistics of coincidence counts and do not take into account possible systematic errors. For detailed discussion of experimental data refer to *SI Appendix, State tomography results*.

maximal violation of the 3D generalization of the Bell inequalities (6, 39). The generation rates of our setup are around 1,200 Hz for the 3D states and around 1,400 Hz for the 2D states. It is important to mention here that these count rates are the actually detected ones. All losses from the detection scheme, such as spatial light modulators, detectors, and other optical elements, are already included. With the constant total pump power the two rates should be equal. The reason why the former is smaller is that an SPDC process is less efficient when pumped by a beam with a nonzero number of OAM quanta, as is the case for crystals B and C. This effect is not present when all crystals are pumped by a fundamental mode as proposed in the scheme in Fig. 1.

The real parts of the density matrices for three of the states presented in Table 1 are displayed in Fig. 3. There, the measurement results (solid bars) are compared to the theoretical expectations (translucent bars). The average fidelity $87.3 \pm 2.2\%$ of three-dimensionally entangled states does not decrease significantly when compared to the average fidelity of $89.8 \pm 0.9\%$ of 2D states. The quality of the entangled states is thus mostly unaffected when going from two to three dimensions and indicates that our approach can be feasible for even higher dimensions. The fidelities reported in Table 1 do not reach unity for two main reasons. First, imperfect coherence of the SPDC processes amounts to roughly 5% decrease in the fidelities for all reported states irrespective of their dimensionality. The main limitation for achieving higher coherence is slight distinguishability of the SPDC sources, which we attribute to small differences in the spectral and polarization degrees of freedom of the down-converted photons. Second, an imprecise setting of local phases, slight misalignment, and the presence of higher-order OAM modes lead to an extra decrease of fidelities, which varies for different states. This explains the range 85 to 90% of fidelities for different 3D states. Nevertheless, none of these imperfections are of fundamental nature. We analyze the causes of these imperfections in detail in *SI Appendix* and therefore facilitate technical improvements in future development iterations. Complete state tomography data are presented in *SI Appendix, State tomography results*.

Alternative Designs

The modular structure of the setup gives rise to the scalability of our scheme in the sense that to increase the entanglement dimension by one requires a mere addition of a single crystal and a single mode shifter (SPP). To further improve the performance, some modifications to our experimental implementation can be made. We adopted the Mach–Zehnder interferometric configuration in our experiment. This gives us freedom to access and manipulate the pump and down-conversion beams separately with no need of custom-made components. The distance between two successive crystals in our current setup is 600 mm. Due to these large interferometers, active stabilization is inevitable. However, scaling down the distances and employing integrated fabrication techniques as used in microchip fabrication lead to significantly more stable interferometers. An alternative approach is to circumvent interferometers completely by, for example, using wavelength-dependent phase shifters and q plates (50, 51), which is inherently stable.

The framework of entanglement by path identity can be easily employed to generate hyperentangled states. Our source of photon pairs can be modified to produce polarization–OAM hyperentanglement when the noncollinear type II SPDC process is utilized in each crystal. This way, the two photons are already created in a polarization-entangled state and due to the geometry of the setup they become also entangled in OAM. In addition, the framework represents a more efficient alternative to traditional techniques to generate multipartite entanglement (19). For instance, in the case of the 3D three-photon Greenberger–Horne–Zeilinger (GHZ) state, the design based on the entanglement by path identity produces entangled states with probability that is eight times larger than when one uses the traditional approach based on interference (12).

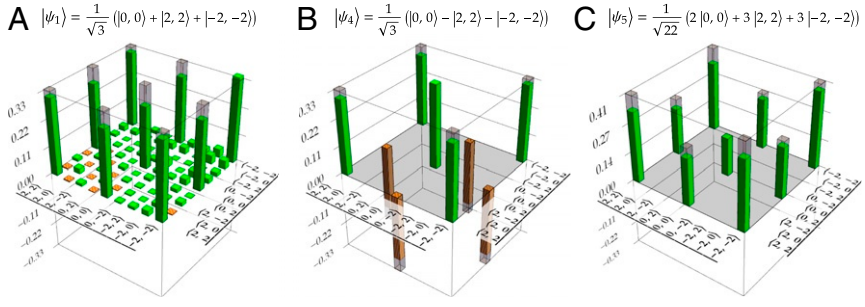


Fig. 3. Examples of the three-dimensionally entangled states $|\psi_1\rangle$, $|\psi_4\rangle$, and $|\psi_5\rangle$ produced (Table 1). With our method we can control the relative phases, as demonstrated in *A* and *B*, as well as relative magnitudes, as shown in *C*. Only real parts are shown; imaginary parts lie in the range $(-0.12, 0.12)$ for all cases. Background noise contributions lie in the range $(-0.04, 0.04)$ for all cases and are explicitly shown only in *A*. This background is omitted in *B* and *C* to improve readability. Green (solid) and orange (hatched) bars represent positive and negative values of reconstructed density matrices, respectively. Gray translucent bars represent the theoretical expectation. Fidelities of the measured states with their reference states are $87.0 \pm 0.5\%$, $89.0 \pm 0.4\%$, and $84.8 \pm 0.8\%$, respectively.

Conclusion

We performed a proof-of-principle experiment of a method called path identity to generate high-dimensionally entangled quantum states. In contrast to previous entanglement creation schemes, here the form of the created quantum state is not dependent on the photon pair creation process itself, but the geometrical arrangement of the setup. Besides its conceptual difference, our approach has two core strengths: a simple and a modular design. The simple geometry-based approach allows us to design an experimental layout that creates versatile and high-dimensionally entangled photon pairs in OAM. These states can be readily utilized in various applications such as superdense coding, high-dimensional quantum teleportation, and violations of generalized Bell inequalities.

We confirmed the modularity of our source by generating different entangled quantum states in two and three dimensions. Thereby we found that the average fidelity of the created states is not decreasing significantly. Thus we believe that extending this modular arrangement is possible and will lead to even higher-dimensionally entangled states in the future. Another very appealing feature of our method is that different families of spatial modes can be used. It is, therefore, possible to create high-dimensional entangled photon pairs in specific modes optimized for free-space communication or even fiber-based systems.

Data Availability. All study data are included in this article and *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by the Austrian Academy of Sciences, the European Research Council (Simulators and Interfaces with Quantum Systems [SIQS] Grant 600645 EU-FP7-ICT), the Austrian Science Fund: F40 (Special Research Programmes [SFB] Foundations and Applications of Quantum Science [FoQuS]) and W 1210-N25 (Complex Quantum Systems [CoQuS]), and the University of Vienna via the project QUESS (Quantum Experiments on Space Scale).

1. C. Wang, F. G. Deng, Y. S. Li, X. S. Liu, G. L. Long, Quantum secure direct communication with high-dimension quantum superdense coding. *Phys. Rev. A* **71**, 044305 (2005).
2. J. T. Barreiro, T. C. Wei, P. G. Kwiat, Beating the channel capacity limit for linear photonic superdense coding. *Nat. Phys.* **4**, 282–286 (2008).
3. X. M. Hu *et al.*, Beating the channel capacity limit for superdense coding with entangled ququarts. *Sci. Adv.* **4**, eaat9304 (2018).
4. M. Huber, M. Pawłowski, Weak randomness in device-independent quantum key distribution and the advantage of using high-dimensional entanglement. *Phys. Rev.* **88**, 032309 (2013).
5. N. J. Cerf, M. Bourennane, A. Karlsson, N. Gisin, Security of quantum key distribution using *d*-level systems. *Phys. Rev. Lett.* **88**, 127902 (2002).
6. D. Collins, N. Gisin, N. Linden, S. Massar, S. Popescu, Bell inequalities for arbitrarily high-dimensional systems. *Phys. Rev. Lett.* **88**, 040404 (2002).
7. J. Ryu, C. Lee, M. Żukowski, J. Lee, Greenberger-Horne-Zeilinger theorem for *N* qudits. *Phys. Rev.* **88**, 042101 (2013).
8. J. Lawrence, Rotational covariance and Greenberger-Horne-Zeilinger theorems for three or more particles of any dimension. *Phys. Rev.* **89**, 012105 (2014).
9. T. Vértesi, S. Pironio, N. Brunner, Closing the detection loophole in Bell experiments using qudits. *Phys. Rev. Lett.* **104**, 060401 (2010).
10. X. L. Wang *et al.*, Quantum teleportation of multiple degrees of freedom of a single photon. *Nature* **518**, 516–519 (2015).
11. M. Malik *et al.*, Multi-photon entanglement in high dimensions. *Nat. Photonics* **10**, 248–252 (2016).
12. M. Erhard, M. Malik, M. Krenn, A. Zeilinger, Experimental Greenberger-Horne-Zeilinger entanglement beyond qubits. *Nat. Photonics* **12**, 759–764 (2018).
13. M. Kues *et al.*, On-chip generation of high-dimensional entangled quantum states and their coherent control. *Nature* **546**, 622–626 (2017).
14. R. T. Thew, A. Acín, H. Zbinden, N. Gisin, Bell-type test of energy-time entangled qutrits. *Phys. Rev. Lett.* **93**, 010503 (2004).
15. H. de Riedmatten *et al.*, Tailoring photonic entanglement in high-dimensional Hilbert spaces. *Phys. Rev. A* **69**, 050304 (2004).
16. A. Tiranov *et al.*, Quantification of multidimensional entanglement stored in a crystal. *Phys. Rev. A* **96**, 040303 (2017).
17. C. Schaeff, R. Polster, M. Huber, S. Ramelow, A. Zeilinger, Experimental access to higher-dimensional entangled quantum systems using integrated optics. *Optica* **2**, 523 (2015).
18. J. Wang *et al.*, Multidimensional quantum entanglement with large-scale integrated optics. *Science* **360**, 285–291 (2018).
19. M. Krenn, A. Hochrainer, M. Lahiri, A. Zeilinger, Entanglement by path identity. *Phys. Rev. Lett.* **118**, 080401 (2017).
20. M. Krenn, X. Gu, A. Zeilinger, Quantum experiments and graphs: Multiparty states as coherent superpositions of perfect matchings. *Phys. Rev. Lett.* **119**, 240403 (2017).
21. M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, A. Zeilinger, Automated search for new quantum experiments. *Phys. Rev. Lett.* **116**, 090405 (2016).
22. L. Allen, M. W. Beijersbergen, R. J. C. Spreeuw, J. P. Woerdman, Orbital angular momentum of light and the transformation of Laguerre-Gaussian laser modes. *Phys. Rev. A* **45**, 8185–8189 (1992).
23. G. Molina-Terriza, J. P. Torres, L. Torner, Twisted photons. *Nat. Phys.* **3**, 305–310 (2007).
24. H. Rubinsztein-Dunlop *et al.*, Roadmap on structured light. *J. Opt.* **19**, 013001 (2017).
25. M. J. Padgtt, Orbital angular momentum 25 years on [Invited]. *Opt. Express* **25**, 11265–11274 (2017).
26. M. Erhard, R. Fickler, M. Krenn, A. Zeilinger, Twisted photons: New quantum perspectives in high dimensions. *Light Sci. Appl.* **7**, 17146 (2018).
27. A. Mair, A. Vaziri, G. Weihs, A. Zeilinger, Entanglement of the orbital angular momentum states of photons. *Nature* **412**, 313–316 (2001).
28. F. M. Miatto, A. M. Yao, S. M. Barnett, Full characterization of the quantum spiral bandwidth of entangled biphotons. *Phys. Rev. A* **83**, 033816 (2011).
29. A. C. Dada, J. Leach, G. S. Buller, M. J. Padgett, E. Andersson, Experimental high-dimensional two-photon entanglement and violations of generalized Bell inequalities. *Nat. Phys.* **7**, 677–680 (2011).
30. J. Romero, D. Giovannini, S. Franke-Arnold, S. M. Barnett, M. J. Padgett, Increasing the dimension in high-dimensional two-photon orbital angular momentum entanglement. *Phys. Rev. A* **86**, 012334 (2012).
31. M. Krenn *et al.*, Generation and confirmation of a (100 × 100)-dimensional entangled quantum system. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6243–6247 (2014).
32. C. H. Bennett, H. J. Bernstein, S. Popescu, B. Schumacher, Concentrating partial entanglement by local operations. *Phys. Rev. A* **53**, 2046–2052 (1996).
33. A. Vaziri, J. W. Pan, T. Jennewein, G. Weihs, A. Zeilinger, Concentration of higher dimensional entanglement: Qutrits of photon orbital angular momentum. *Phys. Rev. Lett.* **91**, 227902 (2003).
34. E. V. Kovlakov, S. S. Straupe, S. P. Kulik, Quantum state engineering with twisted photons via adaptive shaping of the pump beam. *Phys. Rev. A* **98**, 060301 (2018).
35. S. Liu *et al.*, Coherent manipulation of a three-dimensional maximally entangled state. *Phys. Rev. A* **98**, 062316 (2018).
36. K. Mattle, H. Weinfurter, P. G. Kwiat, A. Zeilinger, Dense coding in experimental quantum communication. *Phys. Rev. Lett.* **76**, 4656–4659 (1996).
37. M. Żukowski, A. Zeilinger, M. A. Horne, A. K. Ekert, 'Event-ready-detectors' Bell experiment via entanglement swapping. *Phys. Rev. Lett.* **71**, 4287–4290 (1993).
38. Y. H. Luo *et al.*, Quantum teleportation in high dimensions. *Phys. Rev. Lett.* **123**, 070505 (2019).
39. A. Acín, T. Durt, N. Gisin, J. I. Latorre, Quantum nonlocality in two three-level systems. *Phys. Rev. A* **65**, 052325 (2002).
40. L. J. Wang, X. Y. Zou, L. Mandel, Induced coherence without induced emission. *Phys. Rev. A* **44**, 4614–4622 (1991).
41. T. J. Herzog, J. G. Rarity, H. Weinfurter, A. Zeilinger, Frustrated two-photon creation via interference. *Phys. Rev. Lett.* **72**, 629–632 (1994).
42. H. Weinfurter *et al.*, Frustrated downconversion: Virtual or real photons?. *Ann. N. Y. Acad. Sci.* **755**, 61–72 (1995).
43. L. Hardy, Source of photons with correlated polarisations and correlated directions. *Phys. Lett. A* **161**, 326–328 (1992).
44. P. G. Kwiat, E. Waks, A. G. White, I. Appelbaum, P. H. Eberhard, Ultrabright source of polarization-entangled photons. *Phys. Rev. A* **60**, R773–R776 (1999).
45. H. S. Zhong *et al.*, 12-photon entanglement and scalable scattershot Boson sampling with optimal entangled-photon pairs from parametric down-conversion. *Phys. Rev. Lett.* **121**, 250505 (2018).
46. Y. I. Bogdanov *et al.*, Qutrit state engineering with biphotons. *Phys. Rev. Lett.* **93**, 230503 (2004).
47. Z. Hradil, Quantum-state estimation. *Phys. Rev. A* **55**, R1561–R1564 (1997).
48. R. Fickler *et al.*, Interface between path and orbital angular momentum entanglement for high-dimensional photonic quantum information. *Nat. Commun.* **5**, 4502 (2014).
49. J. Bavaresco *et al.*, Measurements in two bases are sufficient for certifying high-dimensional entanglement. *Nat. Phys.* **14**, 1032–1037 (2018).
50. L. Marrucci, C. Manzo, D. Paparo, Optical spin-to-orbital angular momentum conversion in inhomogeneous anisotropic media. *Phys. Rev. Lett.* **96**, 163905 (2006).
51. L. Yan *et al.*, Q-plate enabled spectrally diverse orbital-angular-momentum conversion for stimulated emission depletion microscopy. *Optica* **2**, 900 (2015).

EPJ Web of Conferences **70**, 00001 (2014)
DOI: 10.1051/epjconf/20147000001

Entanglement, Information, Causality

Gennaro Auletta[1,a]

[1] *University of Cassino, Italy*
Department Letters and Philosophy

Abstract. The paper is divided in two parts. In the first one a summary of the main issues about quantum non–locality is provided. In the second part, the connections with information and causality are considered. In particular, it is shown that a principle of information causality implies that hyper–correlations among experimental settings are not possible but only correlations among possible outcomes. Since a setting is for measuring a particular observable and the eigenbasis of this observable can be considered a code, this means that information codification is a local procedure.

1 EPR

According to EPR, the *correctness* of a theory consists in the degree of agreement between its conclusions and human experience—the objective reality, while its *completeness* is defined as [10]: A theory is complete if every element of objective reality has a counterpart in it. The aim of the EPR article is to show the *incompleteness* of quantum mechanics in the sense of its inability to give a satisfactory explanation of entities which are considered fundamental—in a word, it is a 'disproof' and not a positive proof. Indeed, theories can be disproved by experience and (even thought) experiments.

The core of the argument is constituted by the *Separability principle*, which we can express as follows: Two dynamically independent systems cannot influence each other. The separability principle consists in the assumption that any form of interdependency among physical systems is of dynamical and causal type. Therefore, it is important to carefully distinguish the problem of relativistic locality—i.e., the existence of bounds in the transmission of signals and physical effects—from that of separability, which concerns only the impossibility of a correlation between separated systems in the case in which there are *no dynamical and causal connections*. Part of the EPR argument is that, in the absence of physical interactions, the systems are also separated.

EPR state a sufficient condition for the reality of observables, which can be formulated as follows:

> If, without in any way disturbing a system, we can predict with certainty the value of a physical quantity, then, independently of our measurement procedure, there exists an element of the physical reality corresponding to this physical quantity.

The words "without in any way disturbing a system" tells us that the systems are considered as dynamically independent. The aim of EPR is to show that, assuming separability and the sufficient condition

[a] e-mail: gennaro.auletta@gmail.com

of reality, quantum mechanics is not complete: in logical terms, for quantum mechanics the following statement holds:

$$[(\text{Suff. Cond. Reality}) \wedge (\text{Separability})] \Longrightarrow \neg\text{Completeness}, \tag{1}$$

where $\wedge$, $\neg$, and the arrow are the logical symbols for conjunction (AND), negation and implication, respectively.

The argument of EPR is structured as follows. From (i) the definition of completeness, (ii) the principles of physical reality and separability, and (iii) the fact that, according to quantum mechanics, two non–commuting observables cannot simultaneously have definite values, it follows that the following two statements are incompatible:

- The statement r that the quantum mechanical description of reality given by the wave function is not complete and
- The statement s that when the operators describing two physical quantities do not commute, the two quantities cannot have simultaneous reality.

In formal terms,

$$r \succ\!\prec s, \tag{2}$$

where the symbol $\succ\!\prec$ means a XOR. The meaning of the statement (2) is the following: if it is possible to show that two non–commuting observables have in fact simultaneous reality, we can logically conclude that quantum mechanics cannot be a complete description of reality (from the falsity of s we infer the truth of r).

Let us consider a one-dimensional system $\mathcal{S}$ made of two subsystems $\mathcal{S}_1$ and $\mathcal{S}_2$ interacting during the time interval between t_1 and t_2, with momenta in the position representations:

$$\hat{p}_x^{(1)} = -\imath\hbar\frac{\partial}{\partial x_1} \quad \text{and} \quad \hat{p}_x^{(2)} = -\imath\hbar\frac{\partial}{\partial x_2}, \tag{3}$$

with momentum eigenfunctions

$$\langle p_1 \,|\varphi\rangle = \varphi_p(x_1) \quad \text{and} \quad \langle p_2 \,|\psi\rangle = \psi_p(x_2)\,, \tag{4}$$

respectively. The vectors $|\varphi\rangle$ and $|\psi\rangle$ describe the states of the particles 1 and 2, respectively. The eigenfunctions in the position representation are

$$\varphi_p(x_1) = \frac{1}{\sqrt{2\pi}} e^{\frac{\imath}{\hbar}px_1} \quad \text{and} \quad \psi_p(x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\imath}{\hbar}(x_2 - x_0)p}\,, \tag{5}$$

respectively, where x_0 is a fixed position (constant) and the eigenfunction $\varphi_p(x_1)$ corresponds to eigenvalue $+p$ whilst $\psi_p(x_2)$ corresponds to the eigenvalue $-p$ of the second particle's momentum (in other words, the two particles are moving away from each other with the same direction into opposite senses).

Therefore, the compound system is described by the wave function

$$\Psi(x_1, x_2) = \int_{-\infty}^{+\infty} dp \psi_{-p}(x_2)\varphi_p(x_1). \tag{6}$$

Now, I summarize the scheme of the first thought experiment [2, Chap. 16] [4, Chap. 10]:

(a) We locally measure the momentum on particle 1: let us assume that we find an eigenvalue p'.

(b) Therefore, the state (6) reduces to

$$\psi_{-p'}(x_2)\varphi_{p'}(x_1). \tag{7}$$

(c) Then, it is evident that particle 2 must be in state $\psi_{-p'}$ and this result can be predicted with absolute certainty.

(d) However, we were able to formulate such a prediction without disturbing particle 2 (assumption of separability).

(e) Then, as a consequence of (c) and (d) and of the sufficient condition of reality, $\hat{p}_x^{(2)}$ is an element of reality.

Note that steps (a)–(c) are purely quantum mechanical. Only steps (d)–(e) are connected to the specific EPR argument.

However, if we had chosen to consider another observable of particle 1, say $\hat{x}_1$, whose eigenfunctions are $\varphi_x(x_1)$ (whereas $\psi_x(x_2)$ are the eigenfunctions of the observable $\hat{x}_2$ of particle 2), then we would have written the state Ψ of the compound system as

$$\Psi(x_1, x_2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dx \psi_x(x_2)\varphi_x(x_1). \tag{8}$$

Let us now repeat the previous procedure for the position measurement.

(a') We locally measure the position on particle 1 and find the eigenvalue x'.

(b') Now it is clear that the state (8) reduces to

$$\psi_{x'}(x_2)\varphi_{x'}(x_1). \tag{9}$$

(c') Then, it is evident that the particle 2 must be in the state $\psi_{x'}$ and this result can be predicted with absolute certainty.

(d') However, we have not disturbed particle 2 (assumption of separability).

(e') Then, as a consequence of (c') and (d') and of the sufficient condition of reality, $\hat{x}^{(2)}$ is an element of reality.

Conclusions (e) and (e') look incompatible on the basis of the fact that position and momentum observables of particle 2 do not commute: going back to Propositions *r* and *s* [Eq. (2)], EPR have in this way shown that, assuming that *r* (the quantum mechanical description of reality is not complete) is false, *s* is proved to be false as well since both $\hat{p}_x^{(2)}$ and $\hat{x}^{(2)}$ have simultaneous reality. Then, the previous assumption must be rejected, and *r* must be true. Therefore, according to the EPR argument, quantum mechanics cannot be considered as a complete theory and the wave functions (6) and (8) cannot be considered as complete descriptions of the state of the particles.

2 Bohm's reformulation

The argument as formulate in the original EPR paper is difficult tom test. However, a great step was provided by David Bohm. Consider now two particles with spin $\frac{1}{2}$ that are in a state in which the total spin is zero, that is, they are in a singlet state [8]. They can be produced by a single atom radioactive decay. After a time t_0 the two particles begin to separate and at time t_1 they no longer interact. On the hypothesis that they are not disturbed, the law of angular momentum conservation guarantees that they remain in a singlet state. Considering the projection of the spin along the z–direction, the singlet state may be written in the form

$$|\Psi_0\rangle = \frac{1}{\sqrt{2}}\left(|\uparrow\rangle_1 \otimes |\downarrow\rangle_2 - |\downarrow\rangle_1 \otimes |\uparrow\rangle_2\right), \tag{10}$$

where the subscripts 1 and 2 refer to the particles. This implies that, if a measurement of the spin component along the z direction of particle 1 leads to a result $+1/2$, that of particle 2 along the same direction must give the value $-1/2$, and vice versa. This means that $|\Psi_0\rangle$ is an eigenket of the z component of the spin observables $\hat{\sigma}_{1z}\hat{\sigma}_{2z}$ of the two systems.

Entanglement is a property of the state that is independent of the basis used. In order to see this rotational invariance, let us write it in terms of the z–component eigenvectors as

$$|\Psi_0\rangle = \frac{1}{\sqrt{2}}\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix}_1 \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}_2 - \begin{pmatrix} 0 \\ 1 \end{pmatrix}_1 \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix}_2\right]. \tag{11}$$

Then, $|\Psi_0\rangle$ turns out to be also an eigenvector of $\hat{\sigma}_{1x}\hat{\sigma}_{2x}$ and $\hat{\sigma}_{1y}\hat{\sigma}_{2y}$. For example, let us consider the y orientation. First, let us expand the z–eigenkets into the y–eigenkets:

$$|\uparrow\rangle = \frac{\sqrt{2}}{2}\left(|\uparrow\rangle_y + |\downarrow\rangle_y\right), \tag{12a}$$

$$|\downarrow\rangle = -\frac{\imath\sqrt{2}}{2}\left(|\uparrow\rangle_y - |\downarrow\rangle_y\right). \tag{12b}$$

Then, we can write the singlet state (10) in the y expansion:

$$\begin{aligned}\frac{1}{\sqrt{2}}\left(|\uparrow\rangle_1 \otimes |\downarrow\rangle_2 - |\downarrow\rangle_1 \otimes |\uparrow\rangle_2\right) &= \frac{1}{\sqrt{2}}\left[-\frac{\imath}{2}\left(|\uparrow\rangle_y + |\downarrow\rangle_y\right)_1 \otimes \left(|\uparrow\rangle_y - |\downarrow\rangle_y\right)_2\right. \\ &\quad \left. + \frac{\imath}{2}\left(|\uparrow\rangle_y - |\downarrow\rangle_y\right)_1 \otimes \left(|\uparrow\rangle_y + |\downarrow\rangle_y\right)_2\right] \\ &= \frac{\imath}{\sqrt{2}}\left[\left(|\uparrow\rangle_y\right)_1 \otimes \left(|\downarrow\rangle_y\right)_2 - \left(|\downarrow\rangle_y\right)_1 \otimes \left(|\uparrow\rangle_y\right)_2\right]. \end{aligned} \tag{13}$$

Consequently, we have

$$\left(\hat{\sigma}_{1y}\hat{\sigma}_{2y}\right)|\Psi_0\rangle = \frac{\imath}{\sqrt{2}}\left(\hat{\sigma}_{1y}\hat{\sigma}_{2y}\right)\left[\left(|\uparrow\rangle_y\right)_1 \otimes \left(|\downarrow\rangle_y\right)_2 - \left(|\downarrow\rangle_y\right)_1 \otimes \left(|\uparrow\rangle_y\right)_2\right], \tag{14}$$

which, by making use Pauli matrices and of a reformulation of expression (11) in the y basis, implies

$$\begin{aligned}\left(\hat{\sigma}_{1y}\hat{\sigma}_{2y}\right)|\Psi_0\rangle &= \frac{\imath}{2\sqrt{2}}\begin{bmatrix} 0 & -\imath \\ \imath & 0 \end{bmatrix}_1 \begin{bmatrix} 0 & -\imath \\ \imath & 0 \end{bmatrix}_2 \\ &\quad \times\left[\begin{pmatrix} 1 \\ \imath \end{pmatrix}_1 \otimes \begin{pmatrix} 1 \\ -\imath \end{pmatrix}_2 - \begin{pmatrix} 1 \\ -\imath \end{pmatrix}_1 \otimes \begin{pmatrix} 1 \\ \imath \end{pmatrix}_2\right] \\ &= \frac{\imath}{2\sqrt{2}}\left[\begin{pmatrix} 1 \\ \imath \end{pmatrix}_1 \otimes \begin{pmatrix} -1 \\ \imath \end{pmatrix}_2 - \begin{pmatrix} -1 \\ \imath \end{pmatrix}_1 \otimes \begin{pmatrix} 1 \\ \imath \end{pmatrix}_2\right] \\ &= \frac{\imath}{\sqrt{2}}\left[\left(|\uparrow\rangle_y\right)_1 \otimes \left(-|\downarrow\rangle_y\right)_2 - \left(-|\downarrow\rangle_y\right)_1 \otimes \left(|\uparrow\rangle_y\right)_2\right]. \end{aligned} \tag{15}$$

Now, let us back–substitute this expression into the z expansion:

$$\begin{aligned}\frac{\imath}{\sqrt{2}}\left[\left(|\uparrow\rangle_y\right)_1\left(-|\downarrow\rangle_y\right)_2-\left(-|\downarrow\rangle_y\right)_1\left(|\uparrow\rangle_y\right)_2\right] &= -\frac{\imath}{\sqrt{2}}\frac{1}{2}\left[(|\uparrow\rangle+\imath|\downarrow\rangle)_1(-|\uparrow\rangle+\imath|\downarrow\rangle)_2\right.\\ &\quad\left.-(-|\uparrow\rangle+\imath|\downarrow\rangle)_1(|\uparrow\rangle+\imath|\downarrow\rangle)_2\right]\\ &= \frac{\imath}{\sqrt{2}}(\imath|\uparrow\rangle_1|\downarrow\rangle_2-\imath|\downarrow\rangle_1|\uparrow\rangle_2)\\ &= -\frac{1}{\sqrt{2}}(|\uparrow\rangle_1|\downarrow\rangle_2-|\downarrow\rangle_1|\uparrow\rangle_2)\\ &= -|\Psi_0\rangle, \end{aligned} \tag{16}$$

where I have dropped the symbol $\otimes$ of the sake of simplicity.

3 Bell Theorem

Bell assumed the existence of a hidden parameter λ such that, given λ, the function $A_{\mathbf{a}}$ describing the results obtained by measuring with a device A the spin of the first particle along a chosen direction $\mathbf{a}$ (i.e., the observable $\hat{\boldsymbol{\sigma}}_1\cdot\mathbf{a}$), depends only on λ and on $\mathbf{a}$ [6]. Similarly, the function $B_{\mathbf{b}}$ describing the results when measuring with a device B the spin of the second particle along a chosen direction $\mathbf{b}$ (i.e., $\hat{\boldsymbol{\sigma}}_2\cdot\mathbf{b}$), depends only on $\mathbf{b}$ and λ. The separability principle denies that there can be a form of interdependence between two systems if they do not dynamically interact (factorization rule):

$$A_{\mathbf{a}}B_{\mathbf{b}} = A_{\mathbf{a}}(\lambda)B_{\mathbf{b}}(\lambda), \tag{17}$$

where therefore $A_{\mathbf{a}}$ and $B_{\mathbf{b}}$ represent two deterministic functions of the hidden parameter. Eq. (17) expresses the fact that the probability distributions for the two particles are mutually independent. I assume that the result of each measurement can be either +1 (representing spin up) or -1 (representing spin down), that is,

$$A_{\mathbf{a}}(\lambda) = \pm 1, \quad B_{\mathbf{b}}(\lambda) = \pm 1. \tag{18}$$

Following Eq. (17), if $\wp(\lambda)$ denotes the probability distribution of the hidden parameter λ, then the expectation value of the product of the two components $\hat{\boldsymbol{\sigma}}_1\cdot\mathbf{a}$ and $\hat{\boldsymbol{\sigma}}_2\cdot\mathbf{b}$ is

$$\langle(\hat{\boldsymbol{\sigma}}_1\cdot\mathbf{a})(\hat{\boldsymbol{\sigma}}_2\cdot\mathbf{b})\rangle = \int_\Lambda \wp(\lambda)A_{\mathbf{a}}(\lambda)B_{\mathbf{b}}(\lambda)d\lambda, \tag{19}$$

where Λ represents the set of all possible values of λ.

In the present context, $A_{\mathbf{a}}(\lambda)$ and $B_{\mathbf{b}}(\lambda)$ are functions defining the possible measurement results or the eigenvalues of the measured observables. Since we do not know the values of the hidden parameters λ, we must integrate over all the possible values $\lambda \in \Lambda$. Because $\wp(\lambda)$ is supposed to be a normalized probability distribution, we have

$$\int_\Lambda \wp(\lambda)d\lambda = 1, \tag{20}$$

and, given the values (18), we also have

$$-1 \le \langle\mathbf{a},\mathbf{b}\rangle \le +1, \tag{21}$$

where I have rewritten the expression $\langle(\hat{\boldsymbol{\sigma}}_1\cdot\mathbf{a})\,(\hat{\boldsymbol{\sigma}}_2\cdot\mathbf{b})\rangle$ in the simplified form $\langle\mathbf{a},\mathbf{b}\rangle$. Our aim is to compare the prediction of a deterministic HV theory as expressed by Eq. (19) with the quantum mechanical expectation value, which for the singlet state $|\Psi_0\rangle$ [Eq. (10)] is given by

$$\langle\mathbf{a},\mathbf{b}\rangle_{\Psi_0} = \langle\Psi_0\,|(\hat{\boldsymbol{\sigma}}_1\cdot\mathbf{a})\,(\hat{\boldsymbol{\sigma}}_2\cdot\mathbf{b})|\,\Psi_0\rangle = -\mathbf{a}\cdot\mathbf{b}. \tag{22}$$

This result can be derived when considering the previous products between observables and vectors as sum of Cartesian components

$$\hat{\boldsymbol{\sigma}}_1\cdot\mathbf{a} = a_x\begin{bmatrix}0&1\\1&0\end{bmatrix}_1 + a_y\begin{bmatrix}0&-\imath\\\imath&0\end{bmatrix}_1 + a_z\begin{bmatrix}1&0\\0&-1\end{bmatrix}_1, \tag{23a}$$

$$\hat{\boldsymbol{\sigma}}_2\cdot\mathbf{b} = b_x\begin{bmatrix}0&1\\1&0\end{bmatrix}_2 + b_y\begin{bmatrix}0&-\imath\\\imath&0\end{bmatrix}_2 + b_z\begin{bmatrix}1&0\\0&-1\end{bmatrix}_2, \tag{23b}$$

The expectation value on the singlet state (11) of these two products gives 9 terms, of which the first three have the form

$$\begin{aligned}\langle\Psi_0\,|a_xb_x\hat{\sigma}_{1x}\hat{\sigma}_{2x}|\,\Psi_0\rangle &= \langle\Psi_0|\,\frac{a_xb_x}{\sqrt{2}}\begin{bmatrix}0&1\\1&0\end{bmatrix}_1\begin{bmatrix}0&1\\1&0\end{bmatrix}_2\\ &\quad\times\left[\begin{pmatrix}1\\0\end{pmatrix}_1\otimes\begin{pmatrix}0\\1\end{pmatrix}_2-\begin{pmatrix}0\\1\end{pmatrix}_1\otimes\begin{pmatrix}1\\0\end{pmatrix}_2\right]\\ &= \langle\Psi_0\,|-a_xb_x|\,\Psi_0\rangle = -a_xb_x\,.\end{aligned} \tag{24}$$

Indeed, similar calculations show that we also have

$$\left\langle\Psi_0\left|a_yb_y\hat{\sigma}_{1y}\hat{\sigma}_{2y}\right|\Psi_0\right\rangle = -a_yb_y \quad\text{and}\quad \langle\Psi_0\,|a_zb_z\hat{\sigma}_{1z}\hat{\sigma}_{2z}|\,\Psi_0\rangle = -a_zb_z. \tag{25}$$

The remaining six cross terms are instead all zero, so that we may finally conclude that

$$\langle\Psi_0\,|(\hat{\boldsymbol{\sigma}}_1\cdot\mathbf{a})\,(\hat{\boldsymbol{\sigma}}_2\cdot\mathbf{b})|\,\Psi_0\rangle = -(a_xb_x+a_yb_y+a_zb_z) = -\mathbf{a}\cdot\mathbf{b}. \tag{26}$$

When the two orientations **a** and **b** are parallel, quantum mechanical calculations [see Eq. (15)] show that

$$\langle\mathbf{a},\mathbf{a}\rangle_{\Psi_0} = -1, \tag{27}$$

as it should be since there is a perfect *anticorrelation* (spin–up versus spin–down) between the results of the two measurements.

Since the value given by Eq. (27) for perfect anticorrelation is an experimental fact, also a HV theory must satisfy this requirement. On the other hand, $\langle\mathbf{a},\mathbf{a}\rangle = -1$ holds if and only if we also have

$$A_{\mathbf{a}}(\lambda) = -B_{\mathbf{a}}(\lambda), \tag{28}$$

for any direction **a**. In this case, Eq. (19) reaches the minimum value [see also Eq. (21)]. Under this assumption, we can drop any reference to the *B* device and rewrite Eq. (19) as

$$\langle\mathbf{a},\mathbf{b}\rangle = -\int d\lambda\wp(\lambda)A_{\mathbf{a}}(\lambda)A_{\mathbf{b}}(\lambda). \tag{29}$$

Now we consider two alternative orientations, say **b** and **c**, of the spin measurement of particle 2:

$$\begin{aligned}\langle\mathbf{a},\mathbf{b}\rangle-\langle\mathbf{a},\mathbf{c}\rangle &= -\int d\lambda\wp(\lambda)[A_{\mathbf{a}}(\lambda)A_{\mathbf{b}}(\lambda)-A_{\mathbf{a}}(\lambda)A_{\mathbf{c}}(\lambda)]\\ &= \int d\lambda\wp(\lambda)A_{\mathbf{a}}(\lambda)A_{\mathbf{b}}(\lambda)[A_{\mathbf{b}}(\lambda)A_{\mathbf{c}}(\lambda)-1],\end{aligned} \tag{30}$$

because of the property (18) and since, for any orientation $\mathbf{n}$, we have $[A_{\mathbf{n}}(\lambda)]^2 = 1$, which implies

$$A_{\mathbf{a}}(\lambda)A_{\mathbf{b}}(\lambda)A_{\mathbf{b}}(\lambda)A_{\mathbf{c}}(\lambda) = A_{\mathbf{a}}(\lambda)A_{\mathbf{c}}(\lambda). \tag{31}$$

Then, from Eq. (30) we may prove the inequality

$$|\langle \mathbf{a}, \mathbf{b}\rangle - \langle \mathbf{a}, \mathbf{c}\rangle| \leq \int d\lambda \wp(\lambda)[1 - A_{\mathbf{b}}(\lambda)A_{\mathbf{c}}(\lambda)]. \tag{32}$$

This result is obtained when one considers that for any integrable function $f(x)$, we have

$$\left|\int dx f(x)\right| \quad \leq \quad \int dx\, |f(x)|\,, \tag{33}$$

and, given again the property (18), we also have

$$|A_{\mathbf{b}}(\lambda)A_{\mathbf{c}}(\lambda) - 1| = 1 - A_{\mathbf{b}}(\lambda)A_{\mathbf{c}}(\lambda). \tag{34}$$

Therefore, given the property (20) we finally obtain

$$|\langle \mathbf{a}, \mathbf{b}\rangle - \langle \mathbf{a}, \mathbf{c}\rangle| \ \leq 1 + \langle \mathbf{b}, \mathbf{c}\rangle\,, \tag{35}$$

where

$$\langle \mathbf{b}, \mathbf{c}\rangle = -\int d\lambda \wp(\lambda)A_{\mathbf{b}}(\lambda)A_{\mathbf{c}}(\lambda). \tag{36}$$

A reformulation of the Bell inequality (35) is the so-called CHSH inequality, a widely used form,

$$\left|\langle \mathbf{a}, \mathbf{b}\rangle + \langle \mathbf{a}, \mathbf{b}'\rangle + \langle \mathbf{a}', \mathbf{b}\rangle - \langle \mathbf{a}', \mathbf{b}'\rangle\right| \leq 2, \tag{37}$$

where $\mathbf{a}'$ is a setting alternative to $\mathbf{a}$ as well as $\mathbf{b}'$ to $\mathbf{b}$. We may associate to this inequality the following Bell operator:

$$\begin{aligned}\hat{\mathcal{B}} &= \hat{\boldsymbol{\sigma}}_1 \cdot \mathbf{a}\,(\hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b} + \hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b}') + \hat{\boldsymbol{\sigma}}_1 \cdot \mathbf{a}'\,(\hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b} - \hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b}') \\ &= (\hat{\boldsymbol{\sigma}}_1 \cdot \mathbf{a})\,(\hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b}) + (\hat{\boldsymbol{\sigma}}_1 \cdot \mathbf{a})\,(\hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b}') \\ &\quad + (\hat{\boldsymbol{\sigma}}_1 \cdot \mathbf{a}')\,(\hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b}) + (\hat{\boldsymbol{\sigma}}_1 \cdot \mathbf{a}')\,(\hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b}')\,,\end{aligned} \tag{38}$$

which will play a crucial role later on. I recall indeed that e.g. $\langle \mathbf{a}, \mathbf{b}\rangle$ is a shorthand for $\langle(\hat{\boldsymbol{\sigma}}_1 \cdot \mathbf{a})\,(\hat{\boldsymbol{\sigma}}_2 \cdot \mathbf{b})\rangle$, which allows us to write

$$\left|\left\langle \hat{\mathcal{B}}\right\rangle\right| \leq 2\,. \tag{39}$$

4 Experiments and Loopholes

Tests of the Bell theorem already started in the mid of 1970s. However, several loopholes were discovered that affected these early experiments and could be dealt with step by step. The *first loophole* we consider is the locality loophole. In all experiments, one should consider the possibility that the result of a measurement obtained by using a certain polarizer direction depend on the orientation of the other polarizer. This problem was overcome by Aspect's team [1] as outlined in Fig. 1.

Another difficulty (*second loophole*) concerns the angular correlation: Because of the cosine–squared angular correlation of the directions of the photons emitted in an atomic cascade, an inherent polarization decorrelation is present. Hence the very polarization correlation which could result in a

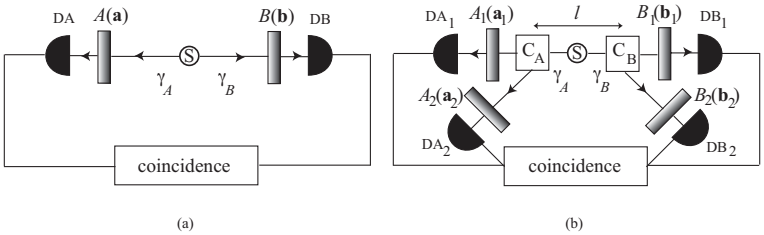


Figure 1. One should consider the possibility that the results obtained using a certain polarization direction could depend on the other polarizer. (a) Friedman–Clauser experiment: The correlated photons γ_A, γ_B coming from the source S impinge upon the linear polarizers A, B oriented in directions $\mathbf{a}, \mathbf{b}$, respectively. (b) Experiment proposed by Aspect: The optical commutator C_A directs the photon γ_A either towards polarizer A_1 with orientation $\mathbf{a}_1$ or to polarizer A_2 with orientation $\mathbf{a}_2$. Similarly for C_B for B_1 and B_2. The two commutators work independently (the time intervals between two commutations are taken to be stochastic). The four joint detection rates are monitored and the orientations $\mathbf{a}_1, \mathbf{a}_2, \mathbf{b}_1, \mathbf{b}_2$ are not changed for the whole experiment. l is the separation between the switches.

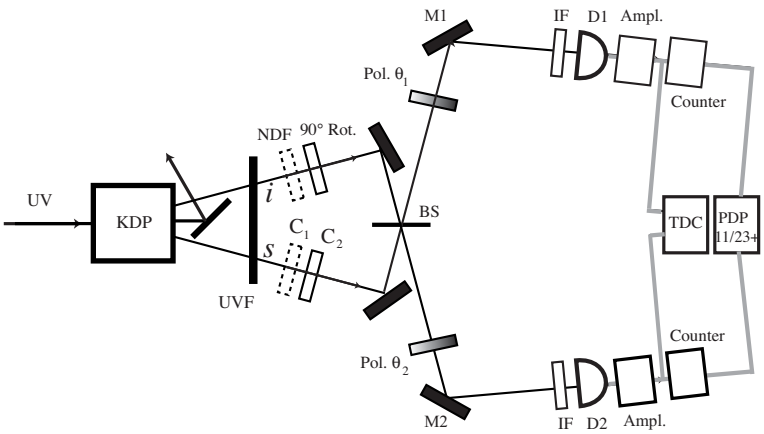


Figure 2. Because of the cosine–squared angular correlation of the directions of the photons emitted in an atomic cascade, an inherent polarization decorrelation is present. Outline of the Alley–Shih and Ou–Mandel's experiment. Light from the 351.1–nm line of an argon–ion laser falls on a non–linear crystal of potassium dihydrogen phosphate (KDP), where down-converted photons of wavelength of about 702 nm are produced. Down–conversion can be tuned in order that linearly polarized signal and idler photons emerge at angles of about ±2° relative to the ultraviolet (UV) pump beam with the electric vector in the plane of the diagram. The *idler* (i) photons pass through a 90° polarization rotator, while the *signal* (s) photons traverse a compensating glass plate C_1 producing an equal time delay. The two photons are then directed from opposite sides towards a beam splitter (BS). The input to the BS consists of an x–polarized s–photon and of a rotated y–polarized i–photon. The light beams emerging from BS, consisting of a mixing of i-photons and s-photons, pass through linear polarizers set at adjustable angles θ_1 and θ_2, through similar interference filters (IF) and finally fall on two photodetectors D_1 and D_2. The photoelectric pulses from D_1 and D_2 are amplified and shaped and fed to the start and stop inputs of a time-to-digital converter (TDC) under computer control which functions as a coincidence counter.

violation of one of the Bell inequalities is reduced for non–collinear photons. The problem can be overcome by using SPDC sources instead of atomic cascade ones [14]. Pairs of photons resulting from SPDC can have an angular correlation of better than 1 mrad, although in general they need not be collinear. The set up is shown in Fig. 2.

A further issue (*third loophole*) is represented by the detection loophole. In fact, we may raise the question of how high the detection efficiencies must be for the experimental confirmation of the

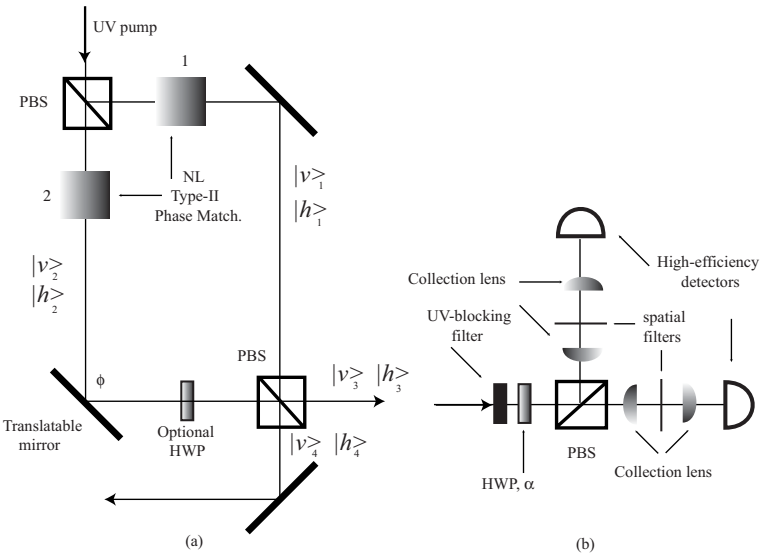


Figure 3. The question is how high the detection efficiency must be for the experimental confirmation of the quantum predictions. In Aspect's experiment the required is 83%. With SPDC experiments, we are obliged to discard part of the counts (when both photons are in the same channel). Proposed experiment for solving the detection loophole. A possible solution is to directly produce a pair of photons in a singlet–kind state avoiding in this way any post–selection.
(a) An ultraviolet pump photon may be spontaneously down-converted in either of two nonlinear crystals, producing a pair of collinear orthogonally polarized photons at half the frequency (type–II phase matching). The outputs are directed toward a second PBS. When the outputs of both crystals are combined with an appropriately relative phase ϕ, a true singlet- or triplet-like state may be produced. By using a half-wave plate (HWP) to effectively exchange the polarizations of photons originating in crystal 2, one overcomes several problems arising from nonideal phase matching. An additional mirror is used to direct the photons into opposite direction towards separated analyzers.
(b) A typical analyzer, including an HWP to rotate by θ the polarization component selected by the analyzing BS, and precision spatial filters to select only conjugate pairs of photons. In an advanced version of the experiment, the HWP could be replaced by an ultrafast polarization rotator (such as Pockels or Kerr cells) to close also the locality loophole.

quantum theoretical predictions. The problem with SPDC–type experiments is that, even with high detection efficiency, one must discard part of the counts, since we are obliged to discard all events where both photons are in the same channel, and one could rise the question whether this selection might represent a bias. Even though this is a remote possibility, in order to exclude any ambiguity a more refined solution is required [12]. A possibility is to directly produce a pair of photons in singlet–type state, thus avoiding any post–selection. One of the first proposals for doing this is shown and summarized in Fig. 3. By means of this apparatus it is possible to produce output photons in the state

$$|\Psi\rangle \simeq |v\rangle_3 \, |h\rangle_4 + e^{\imath\phi} \, (|h\rangle_3 \, |v\rangle_4) \ . \tag{40}$$

5 Non–Locality and Information

I have a general remark. Given any quantum system described by the density matrix $\hat{\rho}$, its von Neumann entropy is [11]

$$S(\hat{\rho}) = -\mathrm{Tr}(\hat{\rho} \ln \hat{\rho}) \ . \tag{41}$$

In fact, the density matrix can be seen as the operator which carries maximal information about the state of the system. If we consider an orthonormal basis $\{|b_k\rangle\}$ of eigenvectors of the density operator

$\hat{\rho}$ for a system $\mathcal{S}$ such that

$$\hat{\rho}\,|b_k\rangle = r_k\,|b_k\rangle \ , \tag{42}$$

where the r_k's are the eigenvalues of $\hat{\rho}$, we may rewrite Eq. (41) as

$$S(\hat{\rho}) = -\sum_j r_j \ln r_j \ , \tag{43}$$

The eigenvectors $|b_k\rangle$ are the possible outcomes of a measurement when we choose to measure the observable of which they are eigenvectors. Let us define the entanglement between systems [5] 1 and 2 as

$$\mathrm{E}(1,2) \quad = \quad S(1,2) - S(1) - S(2) \ , \tag{44}$$

where $S(1,2)$ is the joint (total) entropy of systems 1 and 2, and

$$S(1) = S(\hat{\varrho}_1) \quad \text{and} \quad S(2) = S(\hat{\varrho}_2) \tag{45}$$

are the entropies calculated on the reduced density matrices of the subsystems 1 and 2, respectively, relative to $\hat{\rho}_{12}$. This reflects the fact that entanglement is a quantum form of mutual information: Two entangled systems are correlated because they share an amount of information that is not foreseen classically: indeed the possible outcomes are interdependent.

Are there specific quantum mechanical bounds on information acquisition? Is the bound found with inequality (35) a necessity or are there more rigorous bounds? And if they are, what is their meaning? Let us take advantage of the CHSH inequality (37). Since each of the terms in Eq. (37) lies between -1 and $+1$ [Eq. (21)], the natural upper bound for the entire expression is $+4$. This is precisely the case if we demand that the probabilities satisfy only the causal communication constraint [16], i.e., that they do not violate relativistic locality (what is called non–signaling requirement). In this case, we have

$$\left|\langle \mathbf{a},\mathbf{b}\rangle + \langle \mathbf{a},\mathbf{b}'\rangle + \langle \mathbf{a}',\mathbf{b}\rangle - \langle \mathbf{a}',\mathbf{b}'\rangle\right| \leq 4 \quad \text{or} \quad \left|\left\langle \hat{\mathcal{B}}\right\rangle\right| \leq 4 \ . \tag{46}$$

Indeed, the non–signaling requirement is that the operations one can perform locally here are not influenced by the operations one performs elsewhere, which implies in particular that the probability to obtain a certain outcome (say 1) when choosing the direction $\mathbf{a}$ is independent from the outcomes (either + 1 or -1) when elsewhere one choses a direction $\mathbf{b}$ or $\mathbf{b}'$, that is,

$$\wp_{a,b}(1,1) + \wp_{a,b}(1,-1) = \wp_{a,b'}(1,1) + \wp_{a,b'}(1,-1). \tag{47}$$

Similar considerations hold for any direction. If we consider only this requirement, we are allowed to build the set of probabilities

$$\begin{aligned} \wp_{a,b}(1,1) = \wp_{a,b}(-1,-1) &= \frac{1}{2} \ , \qquad \wp_{a,b'}(1,1) = \wp_{a,b'}(-1,-1) = \frac{1}{2} \\ \wp_{a',b}(1,1) = \wp_{a',b}(-1,-1) &= \frac{1}{2} \ , \qquad \wp_{a',b'}(1,-1) = \wp_{a',b'}(-1,1) = \frac{1}{2}, \end{aligned} \tag{48}$$

while all other probabilities are zero and where I remark that only the $\wp_{a',b'}$ probabilities show anti–correlation.

All the four different expectation values in inequality (46) can be formulated as the following one:

$$\langle \mathbf{a},\mathbf{b}\rangle \quad = \quad \wp_{a,b}(1,1) + \wp_{a,b}(-1,-1) - \wp_{a,b}(1,-1) - \wp_{a,b}(-1,1), \tag{49}$$

where the negative sign of the latter two probabilities is due to the fact that both represent anticorrelations. However, this expectation value in the paramount case in which all probabilities are equal mirrors the separability condition (17), i.e., the absence of correlations ($\langle \mathbf{a}, \mathbf{b} \rangle = 0$) between the two systems. Due to the above assumptions, however, the latter two probabilities are zero so that the whole expression is reduced to

$$\langle \mathbf{a}, \mathbf{b} \rangle = \wp_{a,b}(1,1) + \wp_{a,b}(-1,-1), \tag{50}$$

and similarly for the other three correlations. In this way, taking into account the probabilities (48), the upper bound 4 of inequality (46) is easily obtained.

Let $\hat{O}^a, \hat{O}^{a'}, \hat{O}^b, \hat{O}^{b'}$ be arbitrary Hermitian operators on a Hilbert space $\mathcal{H}$ satisfying the condition $[\hat{O}^a, \hat{O}^b] = 0$ and so on for the other couples $(a, b'), (a', b), (a', b')$ [17]. Moreover, each operator has eigenvalues 1 and -1. We define a generalization of the Bell operator (38), since we are no longer considering the spin observable only,

$$\hat{\mathrm{B}} = \hat{O}^a \hat{O}^b + \hat{O}^{a'} \hat{O}^b + \hat{O}^a \hat{O}^{b'} - \hat{O}^{a'} \hat{O}^{b'}. \tag{51}$$

From the previous assumptions, it follows that the square of each operator is equal to the identity, which implies

$$\begin{aligned} 2\sqrt{2} - \hat{\mathrm{B}} &= \frac{1}{\sqrt{2}} \left[(\hat{O}^a)^2 + (\hat{O}^{a'})^2 + (\hat{O}^b)^2 + (\hat{O}^{b'})^2 \right] - \hat{\mathrm{B}} \\ &= \frac{1}{\sqrt{2}} \left[\left(\hat{O}^a - \frac{\hat{O}^b + \hat{O}^{b'}}{\sqrt{2}} \right)^2 + \left(\hat{O}^{a'} - \frac{\hat{O}^b - \hat{O}^{b'}}{\sqrt{2}} \right)^2 \right] = \hat{\mathrm{A}}. \end{aligned} \tag{52}$$

Now, we wish to prove that $\hat{\mathrm{B}}$ can be expanded as above and that:

$$\left| \left\langle \hat{\mathrm{B}} \right\rangle \right| \leq 2\sqrt{2}\,. \tag{53}$$

Let us expand the operator $\hat{\mathrm{A}}$ as

$$\begin{aligned} &\frac{1}{\sqrt{2}} \left[(\hat{O}^a)^2 + \frac{\left(\hat{O}^b + \hat{O}^{b'} \right)^2}{2} - 2 \frac{\hat{O}^a \left(\hat{O}^b + \hat{O}^{b'} \right)}{\sqrt{2}} + (\hat{O}^{a'})^2 + \frac{\left(\hat{O}^b - \hat{O}^{b'} \right)^2}{2} - 2 \frac{\hat{O}^{a'} \left(\hat{O}^b - \hat{O}^{b'} \right)}{\sqrt{2}} \right] \\ &= \frac{1}{\sqrt{2}} \cdot \frac{1}{2\sqrt{2}} \left[2\sqrt{2}(\hat{O}^a)^2 + \sqrt{2}(\hat{O}^b)^2 + \sqrt{2}(\hat{O}^{b'})^2 + 2\sqrt{2}\hat{O}^b \hat{O}^{b'} - 4\hat{O}^a \hat{O}^b - 4\hat{O}^a \hat{O}^{b'} \right. \\ &\quad \left. + 2\sqrt{2}(\hat{O}^{a'})^2 + \sqrt{2}(\hat{O}^b)^2 + \sqrt{2}(\hat{O}^{b'})^2 - 2\sqrt{2}\hat{O}^b \hat{O}^{b'} - 4\hat{O}^{a'} \hat{O}^b + 4\hat{O}^{a'} \hat{O}^{b'} \right] \\ &= \frac{1}{4} \left[2\sqrt{2}(\hat{O}^a)^2 + \sqrt{2}(\hat{O}^b)^2 + \sqrt{2}(\hat{O}^{b'})^2 + 2\sqrt{2}(\hat{O}^{a'})^2 + \sqrt{2}(\hat{O}^b)^2 + \sqrt{2}(\hat{O}^{b'})^2 \right. \\ &\quad \left. + 2\sqrt{2}\hat{O}^b \hat{O}^{b'} - 2\sqrt{2}\hat{O}^b \hat{O}^{b'} - 4\hat{O}^a \hat{O}^b - 4\hat{O}^a \hat{O}^{b'} - 4\hat{O}^{a'} \hat{O}^b + 4\hat{O}^{a'} \hat{O}^{b'} \right] \\ &= \frac{1}{4} \left[2\sqrt{2} \left((\hat{O}^a)^2 + (\hat{O}^b)^2 + (\hat{O}^{a'})^2 + (\hat{O}^{b'})^2 \right) - 4 \left(\hat{O}^a \hat{O}^b + \hat{O}^{a'} \hat{O}^b + \hat{O}^a \hat{O}^{b'} - \hat{O}^{a'} \hat{O}^{b'} \right) \right] \\ &= \frac{1}{4} \left(8\sqrt{2}\,\hat{I} - 4\hat{\mathrm{B}} \right). \end{aligned} \tag{54}$$

This proves the expansion of $\hat{\mathrm{B}}$. Now, consider that the sum or the difference between Hermitian operators is itself a Hermitian operator, which shows that the following two operatorial expressions are Hermitian:

$$\frac{\hat{O}^b + \hat{O}^{b'}}{\sqrt{2}} \quad \text{and} \quad \frac{\hat{O}^b - \hat{O}^{b'}}{\sqrt{2}}\,. \tag{55}$$

The last step implies that also

$$\hat{O}^a - \frac{\hat{O}^b + \hat{O}^{b'}}{\sqrt{2}} \quad \text{and} \quad \hat{O}^{a'} - \frac{\hat{O}^b - \hat{O}^{b'}}{\sqrt{2}} \tag{56}$$

are. Therefore, the operator

$$\hat{\mathrm{A}} = \frac{1}{\sqrt{2}}\left[\left(\hat{O}^a - \frac{\hat{O}^b + \hat{O}^{b'}}{\sqrt{2}}\right)^2 + \left(\hat{O}^{a'} - \frac{\hat{O}^b - \hat{O}^{b'}}{\sqrt{2}}\right)^2\right], \tag{57}$$

which consists of the sum of squares of Hermitian operators, has clearly an expectation value

$$\left\langle \hat{\mathrm{A}} \right\rangle \geq 0. \tag{58}$$

Since by taking the mean value on both sides of Eq. (53) we have

$$\left\langle \hat{\mathrm{A}} \right\rangle = 2\sqrt{2} - \left\langle \hat{\mathrm{B}} \right\rangle, \tag{59}$$

this leads to the conclusion:

$$\left\langle \hat{\mathrm{B}} \right\rangle \quad \leq \quad 2\sqrt{2}. \tag{60}$$

A similar argument leads to

$$\left\langle \hat{\mathrm{B}} \right\rangle \quad \geq \quad -2\sqrt{2}, \tag{61}$$

which finally implies

$$\left|\left\langle \hat{\mathrm{B}} \right\rangle\right| \leq 2\sqrt{2}. \tag{62}$$

The importance of Tsirelson result lies in the fact that it proves that quantum mechanics *does not* fill the entire gap between the bounds set by Eqs. (37) and (46). The former inequality sets bounds (i.e., 2) for classical separable theories whilst quantum mechanics satisfy the bound $2\sqrt{2}$, which is still stricter than the bound (i.e., 4) imposed by inequality (46). In other words, quantum mechanics certainly allows for correlations that are not allowed by classical HV theories. However, there is a wide spectrum of "hyper–correlations" that do not contradict causal communication constraints (they satisfy the bound imposed by inequality (46)) but are nevertheless not allowed by quantum mechanics (since they do not satisfy inequality (62)). Therefore, we need still to clarify the relations between these different bounds.

To examine this point, let us reformulate the CHSH inequality (37) as an equality with the maximal bound attainable, i.e., B=2, which could be rewritten as [13]

$$\frac{1}{2}\mathrm{B} - 1 = 0, \tag{63}$$

where B expresses again the Bell operator written in terms of a numerical parameter B. However we are interested in more general cases than those allowed by the classical separability requirement. In those case, instead of putting a 0 on the right-hand side we write another numerical parameter, D, as follows:

$$\mathrm{D} = \frac{1}{2}\mathrm{B} - 1. \tag{64}$$

Moreover, we like to write the parameter B as a combination of correlations C_{jk} (where $j,k = a,b,a',b'$) expressed in informational terms, that is, with $j,k = 1,0$. We would also like to express the

possible outputs when Alice and Bob measure in informational term 1,0. In other words, instead of speaking of polarization directions $\mathbf{a}$, $\mathbf{a}'$, $\mathbf{b}$, and $\mathbf{b}'$, or of observables $\hat{O}^a$, $\hat{O}^{a'}$, $\hat{O}^b$, $\hat{O}^{b'}$, we would like to introduce generic inputs $a, b = 0$ and $a', b' = 1$. Moreover, instead to have possible results $-1, 1$, we like to introduce information outputs 0,1. With these assumptions, we rewrite the correlation $\langle \mathbf{a}, \mathbf{b} \rangle$ for the non–signaling case as expressed in the formula (50) as the sum of two conditional probabilities:

$$\mathrm{C}_{00} = \wp(11|00) + \wp(00|00), \tag{65a}$$

where what follows the vertical lines are the inputs and what precedes the vertical line the outputs. Similarly, we have

$$\mathrm{C}_{10} = \wp(11|10) + \wp(00|10), \quad \mathrm{C}_{01} = \wp(11|01) + \wp(00|01), \quad \mathrm{C}_{11} = \wp(10|11) + \wp(01|11),$$

where the first equality is a reformulation of the correlation $\langle \mathbf{a}', \mathbf{b} \rangle$, the second of the correlation $\langle \mathbf{a}, \mathbf{b}' \rangle$, and the latter of the correlation $\langle \mathbf{a}', \mathbf{b}' \rangle$, and again I remark that only the latter one is an anticorrelation.

Then we can write:

$$\begin{aligned} \mathrm{D} &= \frac{1}{2}\mathrm{B} - 1 \\ &= \frac{1}{2}\left(\mathrm{C}_{00} + \mathrm{C}_{01} + \mathrm{C}_{10} - \mathrm{C}_{11}\right) - 1. \end{aligned} \tag{66}$$

Let us now introduce a simplification. In the case in which

$$\mathrm{C}_{00} = \mathrm{C}_{01} = \mathrm{C}_{10} = -\mathrm{C}_{11} \geq 0, \tag{67}$$

we can write

$$\mathrm{C} = \mathrm{C}_{00} = \mathrm{C}_{01} = \mathrm{C}_{10} = -\mathrm{C}_{11}, \tag{68}$$

which implies B= 4C (or C = B/4) that allows us to make the parameter D dependent on C and to rewrite the expression (66) as

$$\mathrm{D}(\mathrm{C}) = \frac{1}{2} \cdot 4\mathrm{C} - 1 \quad = \quad 2\mathrm{C} - 1. \tag{69}$$

It is easy to see that

- When C = 1 we also have D = 1, which implies that $\mathrm{B}_{ns} = 4$, which is precisely the non–signaling case (when only the causal requirement in the transmission of signals is observed).
- Instead, we have the classical separability $\mathrm{B}_c = 2$ when C = 1/2 and D = 0.
- In the quantum case, we have $\mathrm{B}_q = 2\sqrt{2}$ when $\mathrm{C} = 1/\sqrt{2}$ and $\mathrm{D} = \sqrt{2} - 1$.

To have a concrete model, let us briefly consider how teleportation works [7]. The eigenbasis of the Bell operator is given by

$$\left|\Psi^-\right\rangle_{12} = \frac{1}{\sqrt{2}}\left(|\uparrow\rangle_1 |\downarrow\rangle_2 - |\downarrow\rangle_1 |\uparrow\rangle_2\right), \tag{70}$$

$$\left|\Psi^+\right\rangle_{12} = \frac{1}{\sqrt{2}}\left(|\uparrow\rangle_1 |\downarrow\rangle_2 + |\downarrow\rangle_1 |\uparrow\rangle_2\right), \tag{71}$$

$$\left|\Phi^-\right\rangle_{12} = \frac{1}{\sqrt{2}}\left(|\uparrow\rangle_1 |\uparrow\rangle_2 - |\downarrow\rangle_1 |\downarrow\rangle_2\right), \tag{72}$$

$$\left|\Phi^+\right\rangle_{12} = \frac{1}{\sqrt{2}}\left(|\uparrow\rangle_1 |\uparrow\rangle_2 + |\downarrow\rangle_1 |\downarrow\rangle_2\right). \tag{73}$$

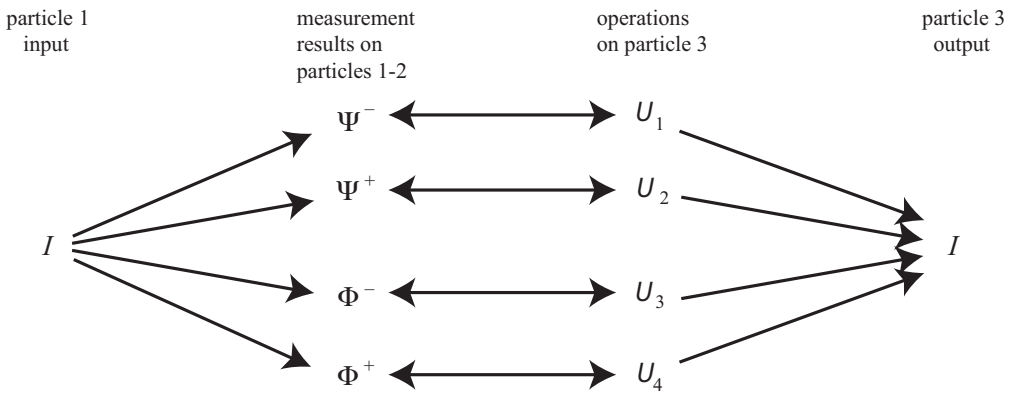


Figure 4. Scheme of teleportation: the fact that each measurement result is mapped in a certain way to the input information I allows through the ebit that Alice classical instructs Bob about the kind of operation to be performed. It is a code.

The state of the three particles can be described as

$$\begin{aligned} |\Psi\rangle_{123} &= \frac{1}{2}\Big[\left|\Psi^-\right\rangle_{12}(-c\,|\uparrow\rangle_3 - c'\,|\downarrow\rangle_3) + \left|\Psi^+\right\rangle_{12}(-c\,|\uparrow\rangle_3 + c'\,|\downarrow\rangle_3) \\ &\quad + \left|\Phi^-\right\rangle_{12}(c\,|\downarrow\rangle_3 + c'\,|\uparrow\rangle_3) + \left|\Phi^+\right\rangle_{12}(c\,|\downarrow\rangle_3 - c'\,|\uparrow\rangle_3)\Big]. \end{aligned} \tag{74}$$

It suffices an unitary operation (a mechanical instruction) to recover the information of Particle 1 on 3 once the Bell operator has been measured:

$$\hat{U}_1 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix};\ \hat{U}_2 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix};\ \hat{U}_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix};\ \hat{U}_4 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{75}$$

We can assume an information causality principle which (in a teleportation protocol) relates to the amount of information that Bob can gain about a data set belonging to Alice, the contents of which are completely unknown to him [15]. Using all his local resources (which may be correlated with her resources) and allowing classical communication from Alice to Bob, the amount of information that the latter can recover is bounded by the information volume (n) of the communication. Namely, if Alice classically communicates n bits to Bob, the total information obtainable by Bob cannot be greater than n. Consider the easiest case in which a two-bit information has been classically transmitted. Then, Bob can at most recover this information (the instruction to perform a particular unitary operation out of four on his particle) and not the whole set of potential information from which Alice has selected the message she has sent. Then, in this simple case, the acquired information must be bound as

$$I \leq 2, \tag{76}$$

since $\lg 4 = 2$. However, if the sole causal (non–signaling) requirement would rule this information exchange, Bob would indeed recover 4 bits (the whole code mapping four outcomes into four operations), thus violating the information causality principle. Therefore, the amazing result that was found by Zukowski *et al.* is that a hypothetical theory which fulfill the requirements of causality but exceeds the Tsirelson bound, also violate the principle of information causality.

Let $\hat{O}_1$ and $\hat{O}_2$ be two observables on subsystems $\mathcal{S}_1$ and $\mathcal{S}_2$ of a system $\mathcal{S}$, respectively, and $\wp(o_a, \mathbf{a}; o_b, \mathbf{b})$ be the probability that the results of a measurement of $\hat{O}_1$ on $\mathcal{S}_1$ and $\hat{O}_2$ on $\mathcal{S}_2$ yield o_a and o_b when certain settings of the measurement apparata are $\mathbf{a}$ and $\mathbf{b}$, respectively. Since this assumption is of absolute generality, we not need to consider the specific spin model previously introduced.

According to Eberhard, the probability distribution of $\hat{O}_1$ (or $\hat{O}_2$), independently of the measurement operations on $\hat{O}_2$ (or $\hat{O}_1$), obtained by integrating or summing the probabilities $\wp(o_a, \mathbf{a}; o_b, \mathbf{b})$ over the possible outcomes o_b (or o_a), needs to be independent of the other setting $\mathbf{b}$ (or $\mathbf{a}$), that is, the two probabilities must depend on local settings only [9]:

$$\sum_{o_b} \wp(o_a, \mathbf{a}; o_b, \mathbf{b}) = \wp(o_a, \mathbf{a}) \; ; \quad \sum_{o_a} \wp(o_a, \mathbf{a}; o_b, \mathbf{b}) = \wp(o_b, \mathbf{b}). \tag{77}$$

According to Eberhard, if this requirement were violated, we would have a *causal non–local interdependence* between the two subsystems. Actually, a violation of the above requirement does not necessarily imply a non–local causal interconnection because there could still be a form of interdependence but satisfying the non–signaling requirement.

In order to prove the theorem, let

$$\hat{P}_{o_a,\mathbf{a}} = |\, o_a, \mathbf{a}\rangle \langle o_a, \mathbf{a}\,| \quad \text{and} \quad \hat{P}_{o_b,\mathbf{b}} = |\, o_b, \mathbf{b}\rangle \langle o_b, \mathbf{b}\,| \tag{78}$$

be the projectors on the state $|o_a, \mathbf{a}\rangle$ of subsystem $\mathcal{S}_1$ when the setting is $\mathbf{a}$ and the outcome $|o_a\rangle$, and on the state $|o_b, \mathbf{b}\rangle$ of subsystem $\mathcal{S}_2$ when the setting is $\mathbf{b}$ and the outcome $|o_b\rangle$, respectively, and $\hat{\rho}$ a density matrix which represents the compound state of $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$. The probability $\wp(o_a, \mathbf{a})$ that, by measuring the observable $\hat{O}_1$ on $\mathcal{S}_1$, we obtain the outcome $|o_a\rangle$ (or the eigenvalue o_a), is

$$\wp(o_a, \mathbf{a}) = \mathrm{Tr}\left[\hat{P}_{o_a,\mathbf{a}}\hat{\rho}\right]. \tag{79}$$

After a measurement of $\hat{O}_1$ when the setting is $\mathbf{a}$ with result o_a we obtain the transformation

$$\hat{\rho} \mapsto \hat{\rho}' = \frac{\hat{P}_{o_a,\mathbf{a}}\hat{\rho}\hat{P}_{o_a,\mathbf{a}}}{\wp(o_a, \mathbf{a})}. \tag{80}$$

If we perform a second measurement on the second subsystem, the conditional probability of obtaining $|o_b\rangle$ (or o_b) by measuring $\hat{O}_2$ when the setting is $\mathbf{b}$, is given by

$$\wp'(o_b, \mathbf{b}|o_a, \mathbf{a}) = \mathrm{Tr}\left[\hat{P}_{o_b,\mathbf{b}}\hat{\rho}'\right] = \frac{\mathrm{Tr}\left[\hat{P}_{o_b,\mathbf{b}}\hat{P}_{o_a,\mathbf{a}}\hat{\rho}\hat{P}_{o_a,\mathbf{a}}\right]}{\wp(o_a, \mathbf{a})}. \tag{81}$$

For any events A and B, the classical probability calculus tells us that the probability of their joint occurrence can be expressed as

$$\wp(A, B) = \wp(A)\wp(B|A). \tag{82}$$

Therefore, the joint probability of obtaining the two results o_a and o_b given the settings $\mathbf{a}$ and $\mathbf{b}$, is given by combining Eqs. (79) and (81):

$$\begin{aligned} \wp(o_a, \mathbf{a}; o_b, \mathbf{b}) &= \wp(o_a, \mathbf{a})\wp'(o_b, \mathbf{b}|o_a, \mathbf{a}) \\ &= \wp(o_a, \mathbf{a}) \frac{\mathrm{Tr}\left[\hat{P}_{o_b,\mathbf{b}}\hat{P}_{o_a,\mathbf{a}}\hat{\rho}\hat{P}_{o_a,\mathbf{a}}\right]}{\wp(o_a, \mathbf{a})} \\ &= \mathrm{Tr}\left[\hat{P}_{o_b,\mathbf{b}}\hat{P}_{o_a,\mathbf{a}}\hat{\rho}\hat{P}_{o_a,\mathbf{a}}\right]. \end{aligned} \tag{83}$$

Given these assumptions, we can obtain the following result that is in accordance with Eqs. (77):

$$\sum_{o_a} \wp(o_a, \mathbf{a}; o_b, \mathbf{b}) = \mathrm{Tr} \sum_{o_a} \left[\hat{P}_{o_b,\mathbf{b}} \hat{P}_{o_a,\mathbf{a}} \hat{\rho} \hat{P}_{o_a,\mathbf{a}} \right] = \mathrm{Tr} \left[\hat{P}_{o_b,\mathbf{b}} \hat{\rho} \right] = \wp(o_b, \mathbf{b}). \tag{84}$$

To derive this result, first note that

$$\sum_{o_a} \mathrm{Tr} \left[\hat{P}_{o_b,\mathbf{b}} \hat{P}_{o_a,\mathbf{a}} \hat{\rho} \hat{P}_{o_a,\mathbf{a}} \right] = \mathrm{Tr} \sum_{o_a} \left[\hat{P}_{o_b,\mathbf{b}} \hat{P}_{o_a,\mathbf{a}} \hat{\rho} \hat{P}_{o_a,\mathbf{a}} \right]. \tag{85}$$

Moreover, I have made use of the cyclic properties of the trace, i.e., given any three arbitrary observables, we have

$$\mathrm{Tr} \left[\hat{O}_1 \hat{O}_2 \hat{O}_3 \right] = \mathrm{Tr} \left[\hat{O}_3 \hat{O}_1 \hat{O}_2 \right] = \mathrm{Tr} \left[\hat{O}_2 \hat{O}_3 \hat{O}_1 \right]. \tag{86}$$

This property implies that

$$\mathrm{Tr} \sum_{o_a} \left[\hat{P}_{o_b,\mathbf{b}} \hat{P}_{o_a,\mathbf{a}} \hat{\rho} \hat{P}_{o_a,\mathbf{a}} \right] = \mathrm{Tr} \sum_{o_a} \left[\hat{P}_{o_a,\mathbf{a}} \hat{P}_{o_b,\mathbf{b}} \hat{P}_{o_a,\mathbf{a}} \hat{\rho} \right]. \tag{87}$$

Moreover, $\hat{P}_{o_a,\mathbf{a}}$ and $\hat{P}_{o_b,\mathbf{b}}$ commute because they pertain to different subsystems, and therefore we have

$$\mathrm{Tr} \sum_{o_a} \left[\hat{P}_{o_a,\mathbf{a}} \hat{P}_{o_b,\mathbf{b}} \hat{P}_{o_a,\mathbf{a}} \hat{\rho} \right] = \mathrm{Tr} \sum_{o_a} \left[\hat{P}_{o_b,\mathbf{b}} \hat{P}_{o_a,\mathbf{a}} \hat{P}_{o_a,\mathbf{a}} \hat{\rho} \right]. \tag{88}$$

However, any orthogonal set of projectors $\{\hat{P}_{o_a,\mathbf{a}}\}$ satisfies the two properties $\hat{P}^2_{o_a,\mathbf{a}} = \hat{P}_{o_a,\mathbf{a}}$ and $\sum_{o_a} \hat{P}_{o_a,\mathbf{a}} = \hat{I}$, from which we finally obtain

$$\mathrm{Tr} \sum_{o_a} \left[\hat{P}_{o_b,\mathbf{b}} \hat{P}_{o_a,\mathbf{a}} \hat{P}_{o_a,\mathbf{a}} \hat{\rho} \right] = \mathrm{Tr} \left[\hat{P}_{o_b,\mathbf{b}} \hat{\rho} \right]. \tag{89}$$

We may proceed in a similar way starting from the conditional probability $\wp'(o_a, \mathbf{a}|o_b, \mathbf{b})$ in order to derive the second equality (77).

What would happen in a world in which the quantum bound is violated but the locality (non–signaling) requirement is satisfied [3]? Let us now reformulate the quantum–mechanical Eqs. (77) in analogy with Eq. (47) as

$$\wp_{a,b}(1,1) + \wp_{a,b}(1,-1) = p_a(1) \quad \text{and} \quad \wp_{a,b}(1,1) + \wp_{a,b}(-1,1) = \wp_b(1), \tag{90}$$

and similarly for the other outcomes. This clearly shows that quantum mechanics requires a full independence of the settings (here expressed by the orientation **a**), which need to be local operations performed in complete separation from other operations that could be performed elsewhere. Quantum correlations are indeed interdependencies of possible outcomes and not of settings. In other words, a violation of the quantum mechanical bound (and of the information causality principle) would imply that there are correlations between settings. If we consider the abstract forms (65) in which I have written the correlations entering in the CHSH inequality, we see that they are expressed in terms of pure conditional probabilities of the form $\wp(11|00)$ Following the customary approach in quantum mechanics (and our physical experience) we have naturally interpreted probabilities of this form as expressing e.g. the probability that both Alice and Bob get the output 1 given that they have both chosen the setting 0. In fact, if Bob knows which was the setting of Alice (whether 0 or 1) he is able to infer which was her outcome. On this procedure is indeed based quantum cryptography.

However, nothing forbids to interpret such a probability as telling us that, in a Bayesian inversion, Bob is able to predict that Alice has chosen the setting 1 once that he knows that he and Alice have obtained the outcome 0. This is still allowed by the non–signaling condition (47). As a matter of fact, given Eqs. (65), Bob is able to predict any setting of Alice if he knows her outcome:

- If Bob choses the setting 0, obtains the outcome 0 and knows that A has also obtained the outcome 0, he knows that she has chosen the setting 0;
- If Bob choses the setting 1, obtains the outcome 0 and knows that A has also obtained the outcome 0, he knows that she has chosen the setting 1;
- If Bob choses the setting 0, obtains the outcome 0 and knows that A has obtained the outcome 1, he knows that she has chosen the setting 0;
- If Bob choses the setting 1, obtains the outcome 0 and knows that A has obtained the outcome 1, he knows that she has chosen the setting 1;
- If Bob choses the setting 0, obtains the outcome 1 and knows that A has obtained the outcome 0, he knows that she has chosen the setting 0;
- If Bob choses the setting 1, obtains the outcome 1 and knows that A has obtained the outcome 0, he knows that she has chosen the setting 1;
- If Bob choses the setting 0, obtains the outcome 1 and knows that A has also obtained the outcome 1, he knows that she has chosen the setting 1;
- If Bob choses the setting 1, obtains the outcome 1 and knows that A has also obtained the outcome 1, he knows that she has chosen the setting 0.

6 Conclusion

In a world in which settings (and not only outcomes) are shared and so there would be kinds of non–local settings, this would imply that also information codification is shared. Indeed, it can be shown that information codification deals with the choice of a basis which in a measurement context is the choice of a setting. In other words, in a world showing hyper–correlations based on the sharing of settings, information codification would be no longer a local procedure.

The fact that this is forbidden justifies quantum information as a general theory of information since

- It satisfies and saturates the bounds that are imposed by the principle of information causality, and in so doing
- It also sets specific constraints on both the possible interdependencies and the possible interactions (also causal interconnections) in our universe.

The most general conclusion is that information can be defined as correlation among possible outcomes or events, as we have seen above for the mutual–information expression of entanglement.

References

[1] Aspect, A. , Dalibard, J., and Roger, G., Physical Review Letters **49**, 1804–1807 (1982).

[2] Auletta, G., Fortunato, M., and Parisi, G., *Quantum Mechanics* (Cambridge University Press, Cambridge, 2009).

[3] Auletta, G., Journal of Modern Physics **2**, 958–61 (2011).

[4] Auletta, G. and Wang, S.–Y., *Quantum Mechanics for Thinkers* (PanStanford Pub., Peking, 2013).

[5] Barnett, S. M. and Phoenix, S. J. D., Physical Review **A40**, 2404–2409 (1989).

[6] Bell, J. S., Physics **1**, 195–200 (1964).

[7] Bennett, C. H., Brassard, G., Crepeau, C., Jozsa, R., Peres, A., and Wootters, W. K., Physical Review Letters **70**, 1895–1899 (1993).

[8] Bohm, D., *Quantum Theory* (Prentice-Hall, New York, 1951).
[9] Nuovo Cimento **46B**, 392–419 (1978).
[10] Einstein, A., Podolsky, B., and Rosen N., Physical Review **47**, 777-780 (1935).
[11] Fano, U., Review of Modern Physics **29**, 74–93 (1957).
[12] Kwiat, P. G., Eberhard, P. H., Steinberg, A. M., and Chiao, R. Y., Physical Review **A49**, 3209–20 (1994).
[13] Masanes, L., Acin, A., and Gisin, N., Physical Review **A73**, 012112-1–9 (2006).
[14] Ou, Z. Y. and Mandel, L., Physical Review Letters **61**, 50–53 (1988).
[15] Pawłowski, M., Paterek, T., Kaszlikowski, D., Scarani, V., Winter, A., Zukowski, M. Z., Nature **461**, 1101–1104 (2009).
[16] Popescu, S. and Rohrlich, D., Foundations of Physics **24**, 379–85 (1994).
[17] Tsirelson, B. S., Letters in Mathematical Physics **4**, 93–100 (1980).

QUANTUM CORRELATIONS AND ENTANGLEMENT

Claude Fabre
Laboratoire Kastler Brossel, Sorbonne Université, ENS, CNRS, Collège de France, Campus Pierre et Marie Curie, 75005 Paris, France
*** claude.fabre@lkb.upmc.fr**

In 1935, Schrödinger introduced the word "entanglement" to describe a situation examined in the famous Einstein-Podolsky-Rosen paper published a few months before. The proper nature of quantum correlations that exist when a two-partite system is in an entangled state was a subject of controversy. In contrast to many other subjects, the debate about the nature of entanglement came quite recently to an end.

https://doi.org/10.1051/photon/202010755

After a period of gestation in the first quarter of the xxth century, Quantum Mechanics, as a comprehensive *ab initio* theory that could be applied to any physical situation, opened up a new era of physics, as it was able to describe in a quantitative and accurate way many systems: atoms, molecules, solids, electromagnetic fields ... Because of this amazing success, there was no doubt among the community of physicists concerning the validity of quantum theory and of its predictions. But this was not the case for the precise understanding of the concepts introduced, which have been, and are still, the object of debate [1]. Schrödinger, as a start, introduced the wavefunction $\psi(\mathbf{r})$ without knowing the exact nature of it. Born postulated that its square gives the probability of presence at point $\mathbf{r}$ and stressed the fundamental stochastic character of the measurement in quantum mechanics. He wrote it in a short footnote at the bottom of his paper [2], and for these few words he was rewarded with the Nobel prize! This in turn raised a wealth of questions: does the wavefunction, and more generally the state vector $|\psi\rangle$, describe a single particle or an ensemble of particles? Is the intrinsic randomness of the measurements a fundamental feature of the quantum world, or the reflection of our present ignorance? These questions, and many others, were the object of intense discussions, in particular between Einstein and Bohr, at the occasion of the Solvay meetings, and contributed to clarify, if not solve, the issues at stake. Following the 1935 "EPR" paper of Einstein, Podolsky and Rosen [3], and the reactions to this paper by Schrödinger [4] and Bohr, ● ● ●

the discussion focussed on the description of two-particle states and on the characterization of correlations between the measurements performed on these particles, their analogies and differences with the classical ones.

Let us take as an example the polarization states of two photons, labeled 1 and 2. We note $|V_1\rangle|V_2\rangle$ the quantum state of two photons of vertical polarization, and $|H_1\rangle|H_2\rangle$ the state of two photons of horizontal polarization. A basic feature of Quantum Mechanics is the superposition principle, which states that any linear combination of quantum states is another bona fide quantum state. It has been popularized by Schrödinger with his famous cat, superposition of a dead cat and an alive cat. We can therefore consider the state

$$|\psi_{12}\rangle = \frac{1}{\sqrt{2}}(|H_1\rangle|H_2\rangle + |V_1\rangle|V_2\rangle)$$

It is easy to show that this state cannot be written as a product of separate polarization states for each photon, so that is not possible to ascribe any polarization state to them separately. The two-photon state must be considered globally. If one measures the polarization of photon 1 using a polarizer of vertical orientation we have 50% probability of finding him with polarization V or H. If we find H for example, the measurement projects the quantum state of photon 1 on state $|H_1\rangle$ and therefore the global state $|\psi_{12}\rangle$ collapses on state $|H_1\rangle|H_2\rangle$. We are therefore sure that the polarization of photon 2 is also H, even when the two photons have been detected very far from each other. The same reasoning is true for a V measurement. There is therefore a perfect correlation between the measurements made on the two photons. For this reason the state $|\psi_{12}\rangle$ is named an entangled state, the english translation of the German word "Verschränklung" introduced by Schrödinger, who coined this property, "not as ONE, but rather as THE characteristic trait of quantum mechanics". This puzzling behaviour does not require any physical link between the two photons, just the application of basic quantum mechanics rules. The argument can be extended to measurements using polarizers of different orientations than vertical and horizontal that also exhibit correlations.

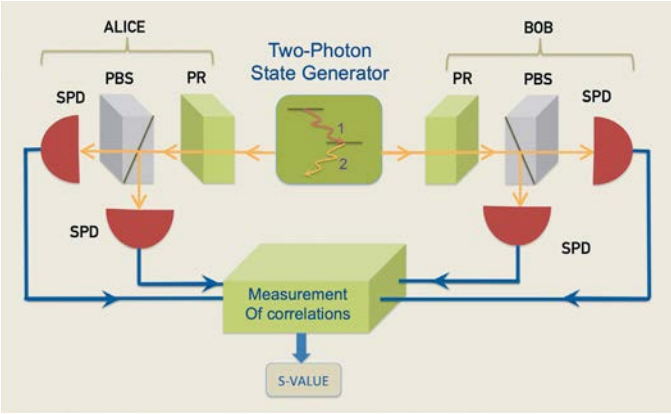


Figure 1: Sketch of an experimental set-up testing Bell inequality on a polarization-entangled two photon state. PR: polarization rotator; PBS: polarizing beamsplitter; SPD: single photon detector.

Of course, correlations do exist also in the classical world. They are even often the basis of scientific approach in many domains of science, for example in sociology, where correlations between apparently unconnected parameters constitute a privileged way to find causal chains. Let us consider the following classical situation: a jeweller, named J, has a stock of earring pairs, 50% in silver, 50% in gold. J randomly choses one pair and sends one earring of the pair to Alice (A) in Australia, and the other to Beatrice (B) in Brazil. When A receives her earring, she finds that it is for example a gold one. She then immediately knows, whatever the distance between them, that B will receive also a gold one. This is a classical case of perfect correlations. Though it seems that some kind of information has been transmitted instantaneously, there is no superluminal effect, because the information that A has on B's jewel is "private". The information is effective and measurable only when A sends a mail to B to communicate him the list of earrings she has received, so that Bob can effectively compare to its known list and measure the correlation.

The classical and quantum situations that we have just described seem at first sight quite similar. This is however not the case: in the classical example, each earring sent by J is made of a definite metal, a "solid" property that is carried all along by the earring. In the quantum example, there is no predetermined value of the spin orientations at the level of the entangled state generation. In addition, the randomness of measurements in the classical example arises from the random choice of earring pairs made by J, not from the probabilistic nature of quantum measurements.

A tempting resolution of the puzzling aspects of the quantum case is to mimic the classical situation by introducing for the two components of each photon pair a "tag" that identifies their common polarization and is carried all the way to Alice and Bob detectors. The value of this tag, named "hidden parameter", is not controlled, so that one has access only to averages over the values of this parameter. This simple picture leads to values of polarization correlations that are identical to the quantum prediction.

The introduction of a supplementary variable implies that the present state of quantum physics is not complete. The possible future mastering of this parameter would eliminate the random character of the quantum measurements. This interpretation of Quantum Mechanics was defended by Einstein (every physicist has in mind his famous statement: "The Old One does not play dice").

In 1964, John Bell made two astonishing discoveries [5]:

- He proved mathematically that the existence of hidden variables is not just a philosophical position. It indeed implies a constraint on measurement results. It showed more precisely that the introduction in the theory of local (i.e. attached to each photon) supplementary variables has indeed a physical consequence: it implies a maximal value of 2 for a well-defined combination, labeled S, of correlations between polarization measurements made by Alice and Bob with two different settings of the polarizer orientations. This is the famous "Bell inequality", which shifted the debate about hidden variables to the domain of experimental physics.
- He also exhibited specific experimental situations for which, according to the ordinary laws of quantum mechanics, one predicts a value of S bigger than 2 (Note that many entangled states do not violate Bell inequality).

Bell's discovery triggered a whole series of experiments [6,7]. In the oldest ones, performed in the 1970's, the entangled state was created by cascaded spontaneous emission on two successive atomic transitions, with two possible paths to the ground state. Most pairs of photons were lost because spontaneous emission is not directional, giving rise to a poor signal to noise ratio. One of the first experiments gave even an unexpected value of S smaller than 2. In experiments performed later in Berkeley and Houston, at the end of the 1970's, the use of laser excitation and improved detection schemes gave S values well above the noise floor. In the beginning of the 1980's the experiment by A. Aspect and coworkers yielded values of S unquestionably above 2, by more than 40 standard deviations. At the end of the 1990's spontaneous parametric down conversion in $\chi^{(2)}$ nonlinear crystal replaced cascades to generate the two-photon polarization entangled states. Phase matching conditions give rise to signal and idler photons emitted in small solid angles, resulting in a significant increase in the quantum efficiency of detection. Nowadays, Bell inequality is strongly violated, and in a short integration time, in photonic systems, but also using two spin 1/2 entangled particles [9].

Such experimentally proven violations of Bell inequality convinced an overwhelming majority of physicists to reject local hidden variables. The debate was actually closed in the 80's, after the Orsay experiments including a fast change of polarization settings during the photons time of flight. However, some theorists raised objections related to the unavoidable imperfections of the experimental protocols. These objections, named "loopholes", are sound from a purely logical point of view, and as such deserve to be examined, but they imply very improbable behaviours of the experimental set-up (a kind of "detector conspiracy") that are very unphysical. Objections concern the possibility of interaction information exchange between Alice and Bob polarization detectors, and a possible "unfair sampling" of the data that were detected, considering the limited collection efficiencies of the photon pairs. Starting from 1981 more and more sophisticated experimental set-ups strived to close this loopholes.

Finally in 2015, the results of 3 "loophole-free" experiments were published [8-11]. Let us briefly describe the one performed by M. Giustina and co-workers in A. Zeilinger's group in the basement of the Hofburg castle in Vienna [10]: the entangled state is generated by spontaneous parametric down conversion in a periodically poled nonlinear crystal and collected in two single mode fibers, at a rate of 3000 pairs per second. While the photons are in flight, fast random number generators choose the two polarization measurement settings. The distances between Alice, Bob, and the entangled state generator, of 30m, are large enough to prevent any kind of causal physical link between them. The detector quantum efficiency is 98%, thanks to the use of TES supraconducting Single Photon Detectors amplified by SQUID. In these optimized experimental conditions, Bell inequality is violated by 11.5 standard deviations on a sample of $3{,}5.10^9$ photon pairs.

After these experiments no serious physicist can now object that the hypothesis of local realistic hidden variables is ruled out by experiments and that an entangled state must be considered as a global, inseparable, entity whatever the distance between its two parties. In addition we must admit that the randomness of Quantum measurements cannot be related to our lack of knowledge about the system. The non-existence of random hidden parameters tells us that it will not be possible to predict for example when exactly an atom will decay by spontaneous emission: the quantum randomness is intrinsic.

Let us finally stress that Bell inequality violating entangled states are not only objects of basic theoretical interest. They are now privileged quantum resources used in applications, such as Quantum Key Distribution and Quantum Teleportation.

RÉFÉRENCES

[1] F. Laloe, Do we really understand Quantum Mechanics?, Cambridge University Press (2019)

[2] M. Born, Zeitschrift für Physik **37**, 863-867, (1926)

[3] A. Einstein, B. Podolsky, N. Rosen, Phys. Rev. **47**, 777 (1935)

[4] E. Schrödinger, Proc. Am. Philos. Soc. **124**, 323-338 (1935)

[5] J. Bell, Physics **1**, 195-200 (1964)

[6] A. Aspect, Bell theorem: a naive view of an experimentalist, in Quantum (un)speakables: from Bell to quantum information, (R. Bertlmann, A. Zeilinger editors, Springer) (2002)

[7] G. Grynberg, A. Aspect, C. Fabre, Polarization-entangled photons and violation of Bell inequality, in Introduction to Quantum Optics, Complement 5C, Cambridge University Press (2010)

[8] J. Miller, Phys. today **69**, 1-14 (2016)

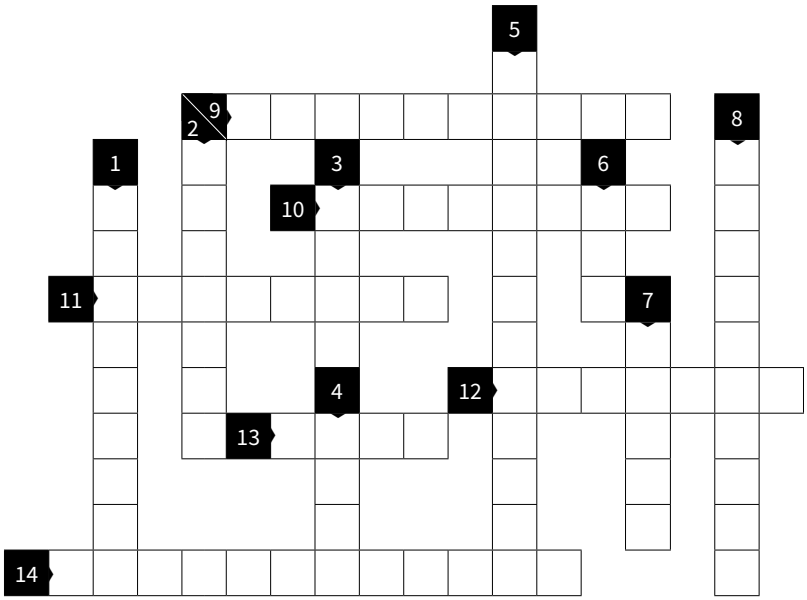
[9] B. Hensen et al., Nature **526**, 682 (2015)

[10] M. Giustina et al., Phys. Rev. Lett. **115**, 250401 (2015)

[11] L. Shalm et al., Phys. Rev. Lett. **115**, 250402 (2015)

CROSSWORDS ON QUANTUM TECHNOLOGIES

SOLUTION ON PHOTONIQUES.COM



1 Property of a quantum operator
2 Big atoms
3 Up or down
4 Inequality
5 Obeys the Schödinger equation
6 Paradox
7 Bob's friend
8 Only with bosons
9 Physical quantity that can be measured
10 States with less uncertainty in one quadrature
11 First condensate
12 Doppler, Sisyphus or evaporative
13 Basic unit in quantum computing
14 Only in quantum mechanics

Journal of Quantum Information Science, 2014, 4, 92-96
Published Online June 2014 in SciRes. http://www.scirp.org/journal/jqis
http://dx.doi.org/10.4236/jqis.2014.42009

Scientific Research

Compatibility of Quantum Entanglement with the Special Theory of Relativity

Burke Ritchie

Lawrence Livermore National Laboratory, Livermore, USA
Email: ritchie@lstc.com

Received 10 March 2014; revised 23 April 2014; accepted 9 May 2014

Abstract

The Einstein-Podolsky-Rosen paradox is resolved dynamically by using spin-dependent quantum trajectories inferred from Dirac's equation for a relativistic electron. The theory provides a practical computational methodology for studying entanglement versus disentanglement for realistic Hamiltonians.

Keywords

Entanglement, Correlation, Spin-Dependent Quantum Trajectory

1. Introduction

The Einstein-Podolsky-Rosen paradox [1] stating that quantum mechanics is incomplete because it violates local realism is resolved by Bell's theorem [2], in which Bell's inequality is violated by quantum mechanics as demonstrated experimentally by Freedman and Clauser [3]. A less abstract way of saying this is the following: although the deterministic description of causality as stated in Einstein's special theory of relativity is violated, nevertheless causality cannot be violated by a physically correct quantum theory of electrons in which local realism is not observed due to the quantum nature of the motion. Such a physically correct theory is the quantum theory of a relativistic electron by Dirac [4] in which the principles of special relativity are incorporated in an equation of motion for a spin-1/2 particle. This statement must be true but yet there is a vagueness or highly abstract character to our understanding of quantum entanglement even after it has been sorted out by the Bell-Freedman-Clauser work. This abstract character, which likely underlies the early perplexity of Einstein, Schroedinger, and others concerning entangled states, exists for two reasons. First electron-electron correlation is not understood in a dynamical sense. We know that two electrons must correlate in space and time, but correlation is understood using time-independent or stationary-state quantum theory both nonrelativistically and relativistically. Second electron spin plays a fundamental role in quantum entanglement, but yet quantum entanglement is understood using Schroedinger theory, in which the electron's spin degree of freedom is absent. In prac-

How to cite this paper: Ritchie, B. (2014) Compatibility of Quantum Entanglement with the Special Theory of Relativity. *Journal of Quantum Information Science*, **4**, 92-96. http://dx.doi.org/10.4236/jqis.2014.42009

tical calculations the omission of spin in the equation describing electrons and their mutual interaction means that the quantum states for two or more electrons must be constructed empirically from experimental observation of how an aggregate of electrons behaves. In other words, there is no mathematical prescription in the many-electron Schroedinger equation itself for Fermi-Dirac statistics, and in fact in many-fermion numerical simulations, *ad hoc* procedures must be used to avert what is called bosonic collapse of the solution. Thus, a physically correct many-electron wave function must be constructed to obey the Pauli Exclusion Principle and Fermi-Dirac statistics. It is well known that a physically correct many-electron wave function in orbital and spin space must be antisymmetric with respect to electron exchange, which is a mathematical recipe to guarantee the Pauli Exclusion Principle and to reconstruct the dynamical information which is otherwise lost in stationary-state theory. The Pauli Exclusion Principle and Fermi-Dirac statistics have only recently been demonstrated on an *ab initio* basis using a dynamical quantum theory of electron exchange-correlation in space and time [5] [6]. In this paper, I show that the previous work [5] [6] also provides an understanding in a dynamical sense of quantum entanglement and disentanglement.

2. Dynamical Theory of Quantum Entanglement

The Pauli exclusion principle, which is fundamental for fermion structure and collision problems, states that each fermion in an ensemble must have a unique set of four quantum numbers three for space and one for spin. For example a pair of electrons can occupy the same spatial orbital only if they have opposite spin states. The canonical examples are the singlet and triplet states (for upper and lower signs respectively) of the helium atom or of the hydrogen molecule,

$$\Psi = \left[\psi_a(1)\psi_b(2) \pm \psi_b(1)\psi_a(2)\right]\left[\alpha\beta \mp \beta\alpha\right], \tag{1}$$

where the arguments refer to the 3-space position vectors of electrons 1 and 2. The second term in square brackets comprises up (alpha) and down (beta) spin states such that the 2-electron spin state is 0 (singlet state) or 1 (triplet state) for upper and lower signs respectively. This point is obvious if the orbitals labeled *a* and *b* are identical such that fermions 1 and 2 occupy the same spatial orbital for the singlet state (upper sign) while the first term in square bracket vanishes for the triplet state (lower sign) since the Pauli exclusion principle is violated for this case.

The singlet state is a canonical example of an entangled state since the two electrons cannot be separated spatially and appear therefore to transfer information between themselves instantaneously. In a way the mysterious nature of the entangled state is illusory due to the incompleteness of the physical theory itself. Firstly the theory is for stationary states with no dynamical information whatsoever between the two correlated electrons. Second Schroedinger theory is spinless such that the 2-electron state written in Equation (1) is an *ad hoc* construction based on experimental observation. It is true that the symmetry of the Schroedinger Hamiltonian with respect to the permutation of electron coordinates allows for pairs of spatial states in the first square bracket to have either even (upper sign) or odd permutation symmetry on the exchange of electrons. Nevertheless the spin dependence of the 2-electron wave function has a totally phenomenological origin such that no *a priori* physical theory exists to explain why the state with total spin angular momentum of 0 is an entangled state while the state with total spin angular momentum of 1 is an unentangled state.

Ironically the paradox is resolved by Einstein's own theory of special relativity. Dirac discovered the correct quantum theory of special relativity for a fermion, which makes it possible to explain both the explicit dynamical nature of entanglement and its dependence on the spin state of a fermion. **Figure 1** and **Figure 2** show 2-electron entanglement and nonentanglement for singlet and triplet states respectively.

The form of Dirac theory which makes this detailed understanding possible is outlined below. First I postulate that a correct dynamical theory for a relativistic electron interacting with other relativistic electrons can be had by replacing the classical relativistic equation of motion for each electron by Dirac's equation as follows [5] [6].

$$\frac{\mathrm{d}\boldsymbol{p}(t)}{\mathrm{d}t} = -\nabla\frac{e^2}{\left|\boldsymbol{r}(t)-\boldsymbol{r}n(t)\right|} \to \mathrm{i}\hbar\frac{\partial\psi_D(\boldsymbol{r},t)}{\partial t} = \left[-\mathrm{i}\hbar c\boldsymbol{\alpha}\cdot\nabla + \beta mc^2 + \frac{e^2}{\left|\boldsymbol{r}-\boldsymbol{s}n(t)\right|}\right]\psi_D(\boldsymbol{r},t) \tag{2}$$

where $\boldsymbol{p} = \gamma m\frac{\mathrm{d}\boldsymbol{r}(t)}{\mathrm{d}t}$, $\gamma = \sqrt{1+\frac{p^2}{m^2c^2}}$, $\psi_D = \begin{pmatrix}\psi\\ \chi\end{pmatrix}$, $\boldsymbol{\alpha} = \begin{pmatrix}0 & \boldsymbol{\sigma}\\ \boldsymbol{\sigma} & 0\end{pmatrix}$, and $\beta = \begin{pmatrix}I & 0\\ 0 & -I\end{pmatrix}$ for Pauli's spin vector $\boldsymbol{\sigma}$ and the 2 × 2 identity matrix I. The generalization to many electrons is obvious. For example for any two

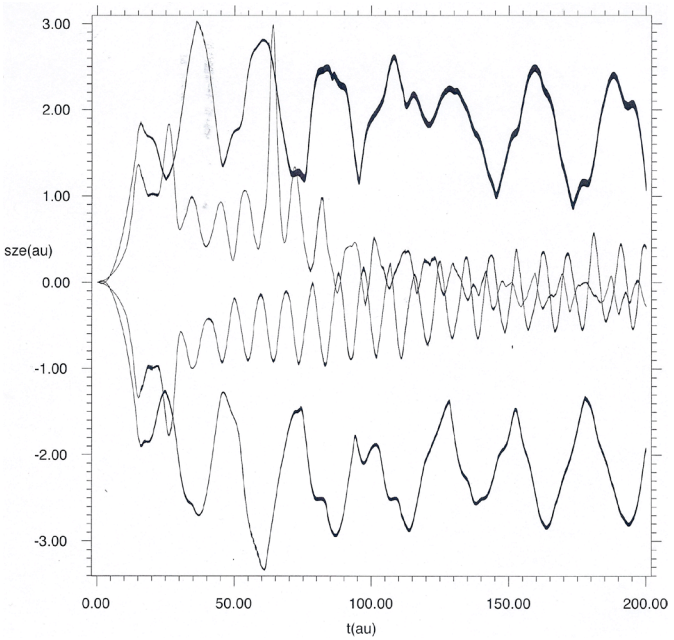


Figure 1. Quantum eigentrajectories in the z direction. Inner curves: R = 1.4 au (R is the internuclear distance). Outer curves: R = 3.0 au. The inner curves show how two electrons with opposite spin states correlate and entangle with increasing time and eventually find the region of covalent bonding located between the two protons fixed at $z = \pm 0.7$ au, while the outer curves show that the electrons remain in the vicinity of the separated atoms for all times. The eigentrajectories are calculated from Equation (5) in the nonrelativistic limit using eigenfunctions for the up and down spin states for the ${}^1\Sigma_g$ state of H_2.

electrons equations of motion analogous to Equation (2) would be written for the primed-variable electron whose interaction with the other electron would now be expressed using the unprimed variable. Notice the passage from classical to quantum dynamics of Coulomb's Law $\frac{e^2}{|\boldsymbol{r}(t)-\boldsymbol{r}n(t)|} \rightarrow \frac{e^2}{|\boldsymbol{r}-\boldsymbol{s}n(t)|}$ for the interaction of any two electrons whose trajectories are at $\boldsymbol{r}(t)$ and $\boldsymbol{r}n(t)$ classically and at $\boldsymbol{r}$ and $\boldsymbol{s}n(t)$ quantum mechanically.

The quantum trajectory is calculated for the unprimed-variable electron as follows. First this electron's velocity field $\boldsymbol{v}(\boldsymbol{r},t)$ is inferred from its current,

$$\boldsymbol{j}(\boldsymbol{r},t) = c\left[\psi^+(\boldsymbol{r},t)\boldsymbol{\sigma}\chi(\boldsymbol{r},t) + \chi^+(\boldsymbol{r},t)\boldsymbol{\sigma}\psi(\boldsymbol{r},t)\right] = \boldsymbol{v}(\boldsymbol{r},t)\rho(\boldsymbol{r},t), \tag{3}$$

where $\rho(\boldsymbol{r},t) = \psi^+(\boldsymbol{r},t)\psi(\boldsymbol{r},t) + \chi^+(\boldsymbol{r},t)\chi(\boldsymbol{r},t)$ and from which a trajectory, $\boldsymbol{s}(t)$, can be calculated from the time integration of the velocity field to find a position field,

$$\boldsymbol{q}(\boldsymbol{r},t) = \int_0^t \mathrm{d}tn\boldsymbol{v}(\boldsymbol{r},tn), \tag{4}$$

and finally by finding the quantum expectation value of the position field,

$$\boldsymbol{s}(t) = \int \mathrm{d}\boldsymbol{r}\left[\psi^+(\boldsymbol{r},t)\boldsymbol{q}(\boldsymbol{r},t)\psi(\boldsymbol{r},t) + \chi^+(\boldsymbol{r},t)\boldsymbol{q}(\boldsymbol{r},t)\chi(\boldsymbol{r},t)\right] \tag{5}$$

and similarly for the primed electron.

In the nonrelativistic regime of electron velocity the current is evaluated in the nonrelativistic limit using $\psi(\boldsymbol{r},t) = \psi_S(\boldsymbol{r},t)\chi_{m_s}$ and $\chi(\boldsymbol{r},t) = \frac{-\mathrm{i}\hbar c\boldsymbol{\sigma}\cdot\nabla\psi(\boldsymbol{r},t)}{E - V + mc^2}$ where

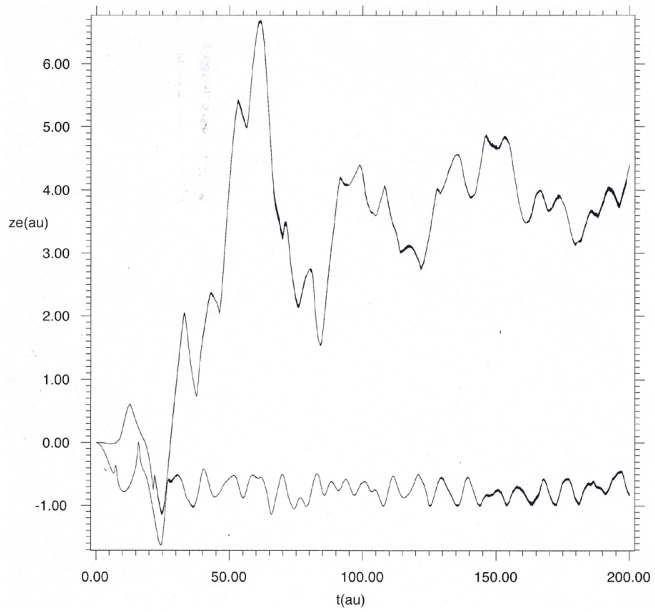


Figure 2. Quantum eigentrajectories showing how the two electrons of H_2 correlate but remain unentangled with increasing time in the formation of an antibonding state. Solid: spin-up electron. Dotted: spin- up electron. The eigentrajectories are calculated from Equation (5) in the nonrelativistic limit using eigenfunctions for two parallel spin states of the $^3\Sigma_u$ state of H_2.

$E - V + mc^2 \cong 2mc^2$ and $\psi_S(\boldsymbol{r},t)$ obeys the time-dependent Schroedinger equation,

$$\mathrm{i}\hbar\frac{\partial\psi_S(\boldsymbol{r},t)}{\partial t} = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{r},t)\right]\psi_S(\boldsymbol{r},t). \tag{6}$$

χ_{m_s} has up (plus sign) or down (minus sign) spin states denoted by $m_s = \pm\frac{1}{2}$ (*i.e.* α or β spin states) respectively. Written out explicitly in terms of the large component the current given by Eq. (3) becomes

$$\boldsymbol{j}(\boldsymbol{r},t) \cong \frac{\hbar}{2m}\left[-\mathrm{i}\left(\psi_S^*\nabla\psi_S - \psi_S\nabla\psi_S^*\right) + \psi_S^*\chi_{m_s}^+(\nabla\times\boldsymbol{\sigma})\psi_S\chi_{m_s} - \psi_S\chi_{m_s}^+(\boldsymbol{\sigma}\times\nabla)\psi_S^*\chi_{m_s}\right], \tag{7}$$

where we have used $\boldsymbol{\sigma}^+ = \boldsymbol{\sigma}$ and the identity, $(\boldsymbol{\sigma}\cdot\boldsymbol{A})(\boldsymbol{\sigma}\cdot\boldsymbol{B}) = \boldsymbol{A}\cdot\boldsymbol{B} + \mathrm{i}\boldsymbol{\sigma}\cdot(\boldsymbol{A}\times\boldsymbol{B})$ from which the identities useful in evaluating the current can be inferred,

$$\boldsymbol{\sigma}(\boldsymbol{\sigma}\cdot\nabla) = \nabla + \mathrm{i}(\nabla\times\boldsymbol{\sigma}) \tag{8a}$$

$$(\boldsymbol{\sigma}\cdot\nabla)\boldsymbol{\sigma} = \nabla + \mathrm{i}(\boldsymbol{\sigma}\times\nabla). \tag{8b}$$

Written out explicitly for up (upper sign) or down (lower sign) spin states

The current in the nonrelativistic regime is

$$\boldsymbol{j}_{nr}(\boldsymbol{r},t) = \frac{\hbar}{m}\left[\mathrm{Im}\,\psi_S^*(\boldsymbol{r},t)\nabla\psi_S(\boldsymbol{r},t) \pm \hat{i}\,\mathrm{Re}\,\psi_S^*(\boldsymbol{r},t)\frac{\partial}{\partial y}\psi_S(\boldsymbol{r},t) \mp \hat{j}\,\mathrm{Re}\,\psi_S^*(\boldsymbol{r},t)\frac{\partial}{\partial x}\psi_S(\boldsymbol{r},t)\right] \tag{9}$$

The first term on the right side of Equation (9), which is independent of spin, is contributed by Schroedinger theory, while the second and third terms are contributed uniquely by Dirac theory. Notice that the current and therefore a quantum trajectory scale like all of the other Schroedinger contributions, namely as c^0 and not as c^{-2}, which have been dropped in the Schroedinger limit of Dirac's equation. It is found in [1] [2] that Pauli's exclusion principle is satisfied automatically on using the spin-dependent quantum trajectories given by Equation (9) to calculate the electron-electron Coulomb potential. Hence one may conclude that electron exchange-correlation—it was recognized by the authors of early highly accurate variational calculations [7] that exchange is au-

tomatically satisfied when correlation is calculated exactly—and Pauli-Dirac statistics are relativistic effects which persist into the nonrelativistic regime. This is obvious on recognizing that spin is a property of a relativistic electron such that in Schroedinger theory the Pauli principle must be satisfied on an *ad hoc* basis from phenomenological observation requiring great mathematical labor to simulate the physical link between electron spin and electron correlation which is omitted in Schroedinger's formulation of quantum theory.

Notice finally that the first-principles understanding of Fermi-Dirac statistics makes available to us a new highly practical computational methodology in which one needs an efficient, accurate solver for the 3D time-dependent Schroedinger equation and an efficient, accurate, energy-conserving integrator for the quantum trajectories. Configuration interaction (CI) calculations are obviated since the time-dependent solution is a superposition of ground and excited states. One should not fuss that the electron-electron Coulomb potential has a mixed evaluation using an independent position variable for one electron and a dependent position variable for the other electron: quantum mechanics allows us latitude to calculate the inverse distance between two point particles as long as it is calculated wave mechanically and not deterministically. The mathematical *bête noir* of conventional time-independent many-electron quantum theory is of course the electron-electron Coulomb potential calculated as an inverse distance using independent position vectors for both electrons. Quantum mechanics does not require us to seek a single wave function for N electrons instead of N wave functions for N electrons, and the former appears to be an accident of the additivity of the Schroedinger Hamiltonian leading to a vast literature on independent-electron approximation methods and on scholastic research on density functionals in which angels are replaced by orbitals. Except for the Bethe-Salpeter equation for two fermions, relativistic invariance is satisfied by a one-body Dirac equation in 4-space: three spatial variables and the scaled time ct. Hence in Dirac theory it is natural to write N wave functions for N fermions as in Equation (2) instead of one wave function for N fermions. As long as the electron-electron potential is written as an exact instantaneous interaction in 3-space and the time, then both electron exchange-correlation and its corollary Fermi-Dirac statistics will be dynamically achieved.

3. Conclusion

In this paper, I have demonstrated quantum entanglement and disentanglement (**Figure 1** and **Figure 2** respectively) in time and space, thereby removing the abstract understanding of these phenomena based on nonrelativistic stationary-state quantum mechanics. This is achieved by inference of a dynamical theory of the electron correlation from Dirac's theory for a relativistic electron such that Fermi-Dirac statistics is obeyed on an *ab initio* basis, thereby elucidating the physical relationship between electron correlation, electron spin, and entangled states.

Acknowledgements

The author is grateful to T. Scott Carman for supporting this work. He is grateful to Professor John Knoblock of the University of Miami for the seminal discussion. This work was performed under the auspices of the Lawrence Livermore National Security, LLC, (LLNS) under Contract No. DE-AC52-07NA27344.

References

[1] Einstein, A., Podolsky, B. and Rosen, N. (1935) Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, **47**, 777. http://dx.doi.org/10.1103/PhysRev.47.777

[2] Bell, J.S. (1964) On the Einstein Podolsky Rosen Paradox. *Physics*, **1**, 195-200.

[3] Freedman, S.J. and Clauser, J.F. (1972) Experimental Test of Local Hidden-Variable Theories. *Physical Review Letters*, **28**, 938. http://dx.doi.org/10.1103/PhysRevLett.28.938

[4] Dirac, P.A.M. (1928) The Quantum Theory of the Electron. *Proceedings of the Royal Society* (*London*), **A117**, 610-624. http://dx.doi.org/10.1098/rspa.1928.0023

[5] Ritchie, B. (2011) Quantum molecular dynamics. *International Journal of Quantum Chemistry*, **111**, 1-7. http://dx.doi.org/10.1002/qua.22371

[6] Ritchie, B. and Weatherford, C.A. (2013) Quantum-Dynamical Theory of Electron Exchange Correlation. *Advances in Physical Chemistry*, 2013, Article ID: 497267. http://dx.doi.org/10.1155/2013/497267

[7] James, H.M. and Coolidge, A.S. (1933) The Ground State of the Hydrogen Molecule. *The Journal of Chemical Physics*, **1**, 825. http://dx.doi.org/10.1063/1.1749252

On Entanglement Assisted Classical Optical Communications

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 24, 2021, accepted March 12, 2021, date of publication March 17, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3066237

On Entanglement Assisted Classical Optical Communications

IVAN B. DJORDJEVIC, (Fellow, IEEE)
Department of Electrical and Computer Engineering, The University of Arizona, Tucson AZ 85721, USA
e-mail: ivan@email.arizona.edu

This work was supported in part by NSF.

ABSTRACT Entanglement assisted communication is advocated by numerous authors as an alternative to classical communication offering significant improvement in channel capacity, in particular in noisy regime. In all those papers it is always assumed that entanglement can be distributed without any imperfections, except for attenuation. We demonstrate that under imperfect pre-shared entanglement distribution, assuming that entanglement distribution channel is modeled as a noisy and lossy Bosonic channel, the entanglement assisted communication can be inferior compared to the classical communication, depending on the parameters of the distribution channel. We identify the conditions under which entanglement assistance can still provide an advantage over the classical case. In particular, when both communication and entanglement distribution channels are not used for entanglement assisted communication but rather for classical transmission instead then the classical capacity is always higher than the entanglement assisted capacity. We also study the entanglement assisted communication under the strong atmospheric turbulence effects.

INDEX TERMS Quantum communication, classical communication, entanglement, entanglement assisted capacity, Holevo capacity, shannon capacity.

I. INTRODUCTION

Quantum information processing (QIP) opens new opportunities for high-precision sensing, secure communications, and ultra-high-performance computing [1]–[3]. Entanglement represents a unique resource for QIP enabling new type of sensors with measurement sensitivities beyond the classical limit, allows quantum computers to solve numerically intractable problems, and provides certifiable security for data transmissions whose security is guaranteed fundamental laws of physics as opposed to unproven mathematical assumptions employed in computational security-based cryptography.

The pre-shared entanglement can also be used, at least in theory, to improve the classical communication capacity [4]–[8]. In addition to secure communications and improved sensor sensitivity, the pre-shared entanglement can be used in distributed quantum computing [9], entanglement assisted (EA) distributed sensing [10], and provably-secure quantum computer access [11], to mention few. The EA classical capacity, that is the maximum of quantum mutual information [4], has been known for decades [4]–[6]; however, the structure of optimum quantum receiver, achieving the EA capacity, has not been determined yet. Moreover, in the determination of EA capacity it is assumed that the distribution of pre-shared entanglement is perfect. To distribute the entangled signal-idler photon pairs either the satellite-to-ground links or fiber-based quantum network, as illustrated in Fig. 1, can be used. Unfortunately, the satellite-to-ground link will experience diffraction loss, atmospheric turbulence effects, scattering effects, and background radiation; and clearly the distribution of entanglement cannot be considered as ideal. In similar fashion, in fiber-based quantum network we cannot ignore the dispersion effects, channel attenuation, and phase noise. Moreover, the entangled states need to be stored in quantum memories before being used and given that the quantum memories are imperfect someone has to use quantum error correction to deal with decoherence effects.

In this paper we study EA channel capacity assuming that the entanglement distribution is imperfect, subject to attenuation and noise, and compare it against classical channel capacity for homodyne and heterodyne detections. We demonstrate that in the presence of imperfect pre-shared entanglement, EA assisted capacity get reduced significantly and when signal-idler photon pairs get distributed over the same noisy channel, used for data transmission, EA assisted

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Masini.

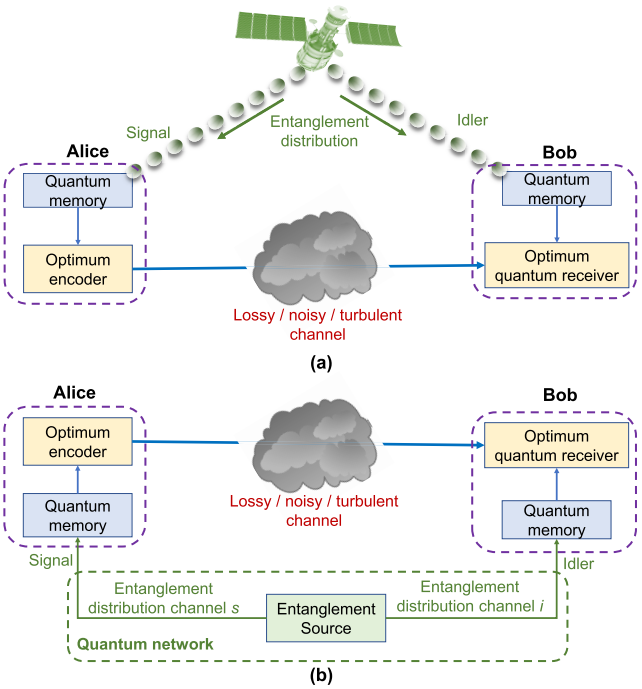


FIGURE 1. Illustrating the entanglement assisted classical optical communication concept: (a) satellite-based entanglement distribution and (b) quantum network-based entanglement distribution.

capacity can be worse than classical capacity with homodyne detection. To achieve the EA capacity it has been shown by Holevo and Werner in [4] that someone needs to use two-mode Gaussian states. We also describe the optimum encoding based on Gaussian modulation of signal photon in two-mode-squeezed-vacuum (TMSV) state. The authors of ref. [8] have shown that EA capacity can also be achieved with random phase modulation. Finally, we study the degradation of EA capacity improvement over classical capacity in the presence of strong atmospheric turbulence effects.

The paper is organized as follows. The realistic entanglement assisted free-space optical (FSO) and fiber-optics communication systems are described in Section II. In Section III, the comparison between EA capacity and classical capacities is performed assuming that entanglement distribution channel is imperfect, modeled as a lossy and noisy Bosonic channel. In Section IV EA communication over FSO links is studied. Concluding remarks are provided in Sec. V.

II. REALISTIC ENTANGLEMENT ASSISTED CLASSICAL OPTICAL COMMUNICATION SYSTEMS

As illustrated in Fig. 1(a), the satellites can be used in entanglement distribution. This is favorable scenario given that satellite-to-ground links are less sensitive to atmospheric turbulence effects compared to ground-to-satellite links. The entangled states are stored in quantum memories and used when needed. Alice employs her signal photon of entangled pair and transmits the classical data, imposed on the signal photon, over lossy, noisy, and possibly turbulent optical channel. On receiver side, Bob employs the entangled idler photon to decide on what was transmitted on signal photon

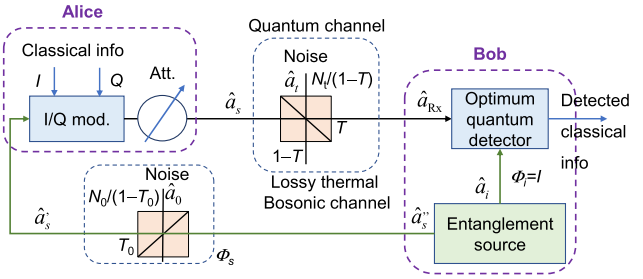


FIGURE 2. Illustrating the entanglement assisted classical communication system model with imperfect pre-shared entanglement distribution. I/Q mod: I/Q modulator, Att.: attenuator (it is optional).

in an optimum quantum receiver. In all papers on entanglement assisted classical communication, such as [4]–[8], it is assumed that the pre-shared entanglement is distributed perfectly, except for the attenuation effect, which is not possible in practice. We consider more realistic scenario as shown in Fig. 2, where the pre-shared entanglement is distributed with the help of two channels: signal channel, denoted by Φ_s, and the idler channel, denoted by Φ_i. We assume that entangled source is located on Bob's side, so that the idler channel is perfect, that is $\Phi_i = I$, where I is the identity operator. On the other hand, the signal channel Φ_s is modelled as a single-mode thermal lossy Bosonic channel, described by the Heisenberg evolution $\hat{a}_s' = \sqrt{T_0}\hat{a}_s'' + \sqrt{1-T_0}\hat{a}_0$, where T_0 is the transmissivity of Bob-to-Alice entanglement distribution channel Φ_s, while $\hat{a}_0$ is a thermal state with the mean photon number being $N_0/(1-T_0)$. The main (Alice-to-Bob) channel is also modelled by the single-mode thermal lossy Bosonic channel, described by the Heisenberg evolution $\hat{a}_{Rx} = \sqrt{T}\hat{a}_s + \sqrt{1-T}\hat{a}_t$, where $T \leq T_0$ is the transmissivity of the main channel, while $\hat{a}_t$ is a thermal state with the mean photon number being $N_t/(1-T)$, $N_t \geq N_0$. Clearly the main channel can also be interpreted as a zero-mean additive white Gaussian noise (AWGN) channel with power-spectral density of N_t and attenuation coefficient of T. Alice modulates the signal mode $\hat{a}_s'$ with the help of an I/Q modulator, as shown in Fig. 2, by effectively performing the following transformation $\hat{a}_s = s\hat{a}_s'$, where $s = s_I + js_Q$ is the transmitted signal constellation point. The coordinates for Gaussian modulation s_I and s_Q are generated from a zero-mean 2-D Gaussian distribution in the digital domain, a digital-to- analog converter (DAC) is used to represent the samples, which are further used as RF inputs of the I/Q modulator. The Gaussian samples are properly scaled to account for I/Q modulator insertion loss such that average number of transmitted signal photons per mode is equal to $N_s = \left\langle \hat{a}_s^\dagger \hat{a}_s \right\rangle = \left\langle s^\dagger s \, (\hat{a}_s')^\dagger \hat{a}_s' \right\rangle$. Alternatively, instead of I/Q modulator, the polar modulator can be used [17]. In related paper [7], in their communication system assisted by two-mode squeezed states authors use an I/Q modulator but in different context, to introduce the displacement to the signal photon state.

From Holevo's papers [4],[5] we know that the quantum limit of classical capacity is given by:

$$C_{\text{without EA}} = g\left(TN_s + N_t\right) - g\left(N_t\right), \tag{1}$$

which is also known as the *Holevo capacity*, wherein $g(x) = (x+1)\log_2(x+1) - x\log_2 x$. Given that according to the uncertainty principle both in-phase and quadrature components of a Gaussian state cannot be simultaneously measured with the complete precision, for homodyne detection the information is encoded on a single quadrature so that the average number of received photon is $4TN_s$, while the average number of noise photons is $2N_t+1$, and the corresponding classical capacity for homodyne detection is $C_{\text{hom}} = 0.5\log_2[1+4TN_s/(2N_t+1)]$. On the other hand, in heterodyne detection both quadratures are used so that the average number of received signal photons will be 0.5*0.5*4 $TN_s = TN_s$ (one-half comes from splitting to two quadratures and second half from heterodyne splitting), while the average number of noise photons per quadrature is $(2N_t+1)/2+1/2 = N_t+1$. The corresponding heterodyne channel capacity will be $C_{\text{het}} = \log_2[1+TN_s/(N_t+1)]$. To achieve the channel capacity in classical case we need to use the Gaussian modulation (GM) by generating samples from two Gaussian sources and with the help of an arbitrary waveform generator (AWG) impose them on the optical carrier by using an I/Q modulator. To achieve the Holevo capacity we need to use the Gaussian state. For instance the coherent state with GM can achieve the Holevo capacity.

The TMSV state, achieving the EA capacity according to [4], can be represented as:

$$\left|TMSV(N_s)\right\rangle_{s,i} = \frac{1}{\sqrt{N_s+1}}\sum_{n=0}^{\infty}\left(\frac{N_s}{N_s+1}\right)^{n/2}|n\rangle_s|n\rangle_i \quad (2)$$

and has the covariance matrix:

$$\mathbf{\Sigma}_{TMSV} = \begin{bmatrix} (2N_s+1)\mathbf{1} & 2\sqrt{N_s(N_s+1)}\mathbf{Z} \\ 2\sqrt{N_s(N_s+1)}\mathbf{Z} & (2N_s+1)\mathbf{1} \end{bmatrix}, \quad (3)$$

where $\mathbf{1}$ is the identity matrix and $\mathbf{Z} = \text{diag}(1,-1)$ is the Pauli Z-matrix.

Given that the action of the beam splitter (BS) can be represented by $BS(\tau) = \begin{bmatrix} \sqrt{\tau}\mathbf{1} & \sqrt{1-\tau}\mathbf{1} \\ -\sqrt{1-\tau}\mathbf{1} & \sqrt{\tau}\mathbf{1} \end{bmatrix}$, in order to determine the covariance matrix after the beam splitter in entanglement distribution channel (see Fig. 2) we need to apply the sympletic operation [1],[3] by

$$BS_c(T_0) = \mathbf{1} \oplus BS(T_0) = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sqrt{T_0}\mathbf{1} & \sqrt{1-T_0}\mathbf{1} \\ \mathbf{0} & -\sqrt{1-T_0}\mathbf{1} & \sqrt{T_0}\mathbf{1} \end{bmatrix} \quad (4)$$

on the input covariance matrix $\mathbf{\Sigma}_{TMSV}$ to obtain:

$$\mathbf{\Sigma}' = BS_c(T_0)\begin{bmatrix} \mathbf{\Sigma}_{TMSV} & \mathbf{0} \\ \mathbf{0} & \sigma'^2\mathbf{1} \end{bmatrix}[BS_c(T_0)]^T, \quad (5)$$

where the variance of thermal state is $N_0/(1-T_0)$ thermal photons. By keeping Alice and Bob submatrices we obtain:

$$\mathbf{\Sigma}'_{AB} = \begin{bmatrix} (2N_s+1)\mathbf{1} & 2\sqrt{T_0N_s(N_s+1)}\mathbf{Z} \\ 2\sqrt{T_0N_s(N_s+1)}\mathbf{Z} & (2N_s'+1)\mathbf{1} \end{bmatrix}, \quad (6)$$

where $N_s' = N_sT_0 + N_0$. By repeating the similar procedure for lossy and noisy Bosonic main channel we obtain the following covariance matrix for the zero-mean Gaussian state $\hat{\rho}_{Rx,i}$:

$$\mathbf{\Sigma}_{AB} = \begin{bmatrix} (2N_s+1)\mathbf{1} & 2\sqrt{T_0TN_s(N_s+1)}\mathbf{Z} \\ 2\sqrt{T_0TN_s(N_s+1)}\mathbf{Z} & (2N_s'+1)\mathbf{1} \end{bmatrix}, \quad (7)$$

where $N_s' = (N_sT_0+N_0)T + N_t$. Clearly, this covariance matrix has the standard form [12]–[16] $\mathbf{\Sigma} = \begin{bmatrix} a\mathbf{1} & \mathbf{C} \\ \mathbf{C} & b\mathbf{1} \end{bmatrix}$ with $a = 2N_s+1$, $b = 2N_s'+1$, $\mathbf{C} = c\mathbf{Z}$, $c = 2\sqrt{T_0TN_s(N_s+1)}$, and the symplectic eigenvalues are given by:

$$\begin{aligned} \nu_{\mp} &= \left[\sqrt{(a+b)^2-4c^2} \mp (b-a)\right]/2 \\ &= \sqrt{(N_s+N_s'+1)^2 - 4T_0TN_s(N_s+1)} \mp (N_s'-N_s). \end{aligned} \quad (8)$$

The corresponding expression for the entanglement assisted channel capacity is now simply:

$$C_{EA} = g(N_s) + g(N_s') - \left[g\left(\frac{\nu_+-1}{2}\right) + g\left(\frac{\nu_--1}{2}\right)\right]. \quad (9)$$

Here we propose to use TMSV state and modulate the signal photon by the zero-mean GM as illustrated in Fig. 2, by effectively mapping $\hat{a}_s' \to \hat{a}_s = s\hat{a}_s'$, where $s = s_I + js_Q$ is the transmitted signal constellation point with s_I and s_Q being generated from the 2-D zero-mean circular Gaussian noise source. The variance of 2-D Gaussian distribution is properly chosen such that the average number of transmitted signal photons is $\left\langle s^\dagger s\,(\hat{a}_s')^\dagger\hat{a}_s'\right\rangle = N_s$ so that the covariance matrix (7) is not affected. In incoming section, we compare the EA classical capacity against both (homodyne and heterodyne) classical and Holevo capacities.

III. ENTANGLEMENT ASSISTED COMMUNICATION VS. CLASSICAL OPTICAL COMMUNICATION

When the pre-shared entanglement can be perfectly distributed, as shown in Fig. 3, the EA capacity can indeed significantly outperform the Holevo capacity in very noisy regime $N_s \ll N_t$. When Alice employs the AWG to perform the GM of the signal photon, she is limited by the finite (l,k) precision (l: number of integer bits plus sign bit, k: number of decimal bits), and there is the limit for the lowest possible N_s. To solve for this problem we can scale the (I,Q) constellation point with a positive number and then apply an attenuator after the I/Q modulator. This scaling number is N_s-dependent. On such a way the (1,5) precision is sufficient to achieve the EA classical capacity. Notice that in these calculations we assume that the optimum quantum receiver is used.

For the realistic scenario, when the pre-shared entanglement is distributed over lossy thermal Bosonic channel, as shown in Fig. 4 the EA capacity improvement over both Holevo and classical capacities get significantly reduced. The Holevo capacity is identical to the homodyne capacity, while the heterodyne capacity is a little bit worse. When the entanglement distribution channel is identical to the communication channel the EA capacity is actually lower than the Holevo/homodyne capacity. When transmissivity of

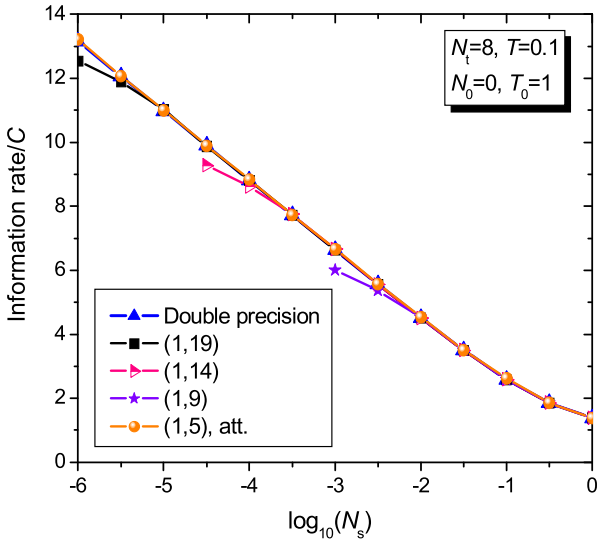


FIGURE 3. The normalized information rate vs. average number of signal photons N_S assuming that the pre-shared entanglement is perfectly distributed ($T_0 = 1$, $N_0 = 0$) and assuming that the optimum quantum receiver is used. The main channel transmissivity is set to $T = 0.1$ and the number of thermal photons to $N_t = 8$.

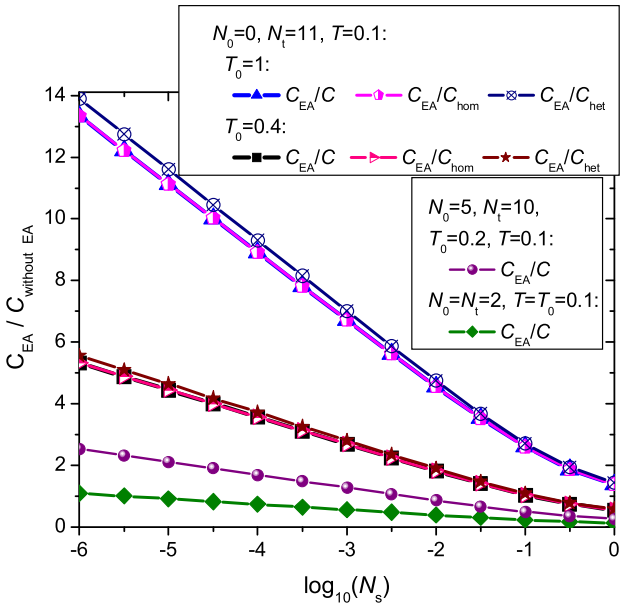


FIGURE 4. The improvement in capacity by entanglement assistance for imperfect pre-shared entanglement distribution.

entanglement distribution channel is higher than the transmissivity of the main (communication) channel, while at the same time is less noisy then the improvement of EA capacity over homodyne capacity is moderate to small (depending on actual value of the average number of signal photons). This particular case is applicable to Fig. 1(a), where the satellite links are used to distribute the pre-shared entanglement, while the free-space optical link with horizontal atmospheric turbulence for transmission of classical information. Namely, in the satellite-to-ground link turbulence is much weaker than the turbulence in horizontal links.

To see the actual improvement, expressed in bits/s/Hz, for this case we provide in Fig. 5 the plots for EA, Holevo,

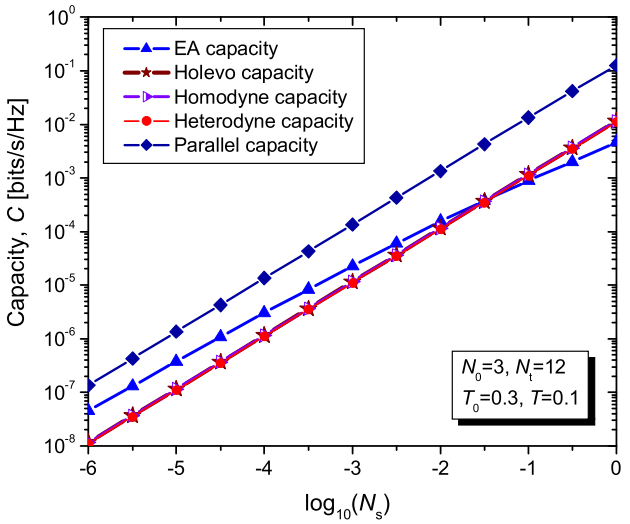


FIGURE 5. EA capacity vs. classical capacity when both main and entanglement distribution channels are noisy.

homodyne, and heterodyne capacities assuming that $N_0 = 3$, $N_t = 12$, $T_0 = 0.3$, and $T = 0.1$. For $N_s = 10^{-6}$, even though that the EA capacity is 3.9 times higher than homodyne capacity, the actual value of EA capacity is only $4.52 \cdot 10^{-8}$ bits/s/Hz. For average number of signal photons N_s between 0.01 and 0.1 the EA capacity performs comparable to the homodyne capacity, while for $N_s > 0.1$, the homodyne capacity outperforms the EA capacity.

In EA communication the entangled pair of photons is used, the signal photon for classical information transmission and the idler photon for entanglement assistance. In classical communication, only the main channel is used, and someone may speculate that this not a fair comparison. So naturally arises the question, what if we use the idler (auxiliary) channel for classical transmission instead? Overall classical capacity will be then addition of capacities of main and axillary channels, which is also shown in Fig. 5 as parallel capacity, which clearly outperforms EA capacity for all N_s values.

IV. ENTANGLEMENT ASSISTED COMMUNICATION OVER FSO LINKS

The beam propagation over FSO links is affected by various effects including the diffraction, atmospheric turbulence, and Mie scattering effects. The attenuation due to diffraction and scattering effects can be modelled by the transmissivity $T \in (0, 1]$ in the main (Alice-to-Bob) channel (see Fig. 2), while atmospheric turbulence is caused by variations in the refractive index of the transmission medium due to spatial variations in temperature and pressure (related to the wind and solar heating [17], [18]).

The atmospheric turbulence can be modelled as the multiplicative noise, described by the following probability density function of irradiance I [17], [18]:

$$f(I) = \frac{2(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)} I^{\frac{\alpha+\beta}{2}-1} K_{\alpha-\beta}\left(2\sqrt{\alpha\beta I}\right), \quad (10)$$

where α and β are the atmospheric turbulence parameters which for zero inner scale are defined, respectively, as [17], [18]:

$$\alpha = \left\{ \exp\left[\frac{0.49\sigma_R^2}{\left(1+1.11\sigma_R^{12/5}\right)^{7/6}} \right] - 1 \right\}^{-1}, \quad \beta = \left\{ \exp\left[\frac{0.51\sigma_R^2}{\left(1+0.69\sigma_R^{12/5}\right)^{5/6}} \right] - 1 \right\}^{-1}, \tag{11}$$

wherein σ_R^2 denotes the Rytov variance, defined as:

$$\sigma_R^2 = 1.23\, C_n^2\, k^{7/6} L^{11/6}. \tag{12}$$

In (12) C_n^2 denotes the refractive structure parameter, k is the wave number ($k = 2\pi/\lambda$, with λ being the wavelength), and L denotes the propagation distance. The Rytov variance can serve an indicator of the strength of the atmospheric turbulence effects. When $\sigma_R^2 < 1$ we say that turbulence is weak, for $\sigma_R^2 \approx 1$ we say that turbulence is medium, the strong turbulence fluctuations are specified with $\sigma_R^2 > 1$, while the saturation regime is defined by $\sigma_R^2 \to \infty$ [17], [18].

The covariance matrix of the zero-mean Gaussian state $\hat{\rho}_{Rx,i}$ in the presence of turbulence can be represented by:

$$\mathbf{\Sigma}_{AB}(I) = \begin{bmatrix} (2N_s+1)\,\mathbf{1} & 2\sqrt{IT_0TN_s(N_s+1)}\mathbf{Z} \\ 2\sqrt{IT_0TN_s(N_s+1)}\mathbf{Z} & \left(2N_s'+1\right)\mathbf{1} \end{bmatrix}, \tag{13}$$

where $N_s' = (N_sT_0 + N_0)\,TI + N_t$. The corresponding channel capacities are now functions of irradiance I, that is $C(I)$ is now a random variable. For instance the EA capacity in the presence of turbulence is evaluated by:

$$C_{EA} = \int_0^\infty C_{EA}(I) f(I)\, dI. \tag{14}$$

To evaluate the *ergodic* capacities we perform the Monte Carlo integration. For every channel use we generate a different realization of irradiance form the gamma-gamma distribution, calculate channel capacities for each realization $C(I)$, and average them out by $C = [\sum C(I)]/L$, where L is the number of realizations. Atmospheric turbulence also introduces the random phase shift, and we assume that a referent classical beam at sufficiently different wavelength is used to estimate and compensate for random phased shift and also to synchronize the transmitter and receiver. This approach has been applied to both classical and EA communications. To compensate for intensity fluctuations someone may use the adaptive optics (AO) approaches [20], [21]. However, the AO can fully compensate the turbulence effects only in weak turbulence regime.

Let us now study how the improvement in EA capacity over classical capacities is affected by atmospheric turbulence effects. For Rytov standard deviation $\sigma_R = 4$ (Rytov variance $\sigma_R^2 = 16$), corresponding to strong turbulence in the main channel, in Fig. 6(a) we summarize the EA capacity improvement over Holevo capacity for imperfect distribution of entanglement, assuming that entanglement distribution

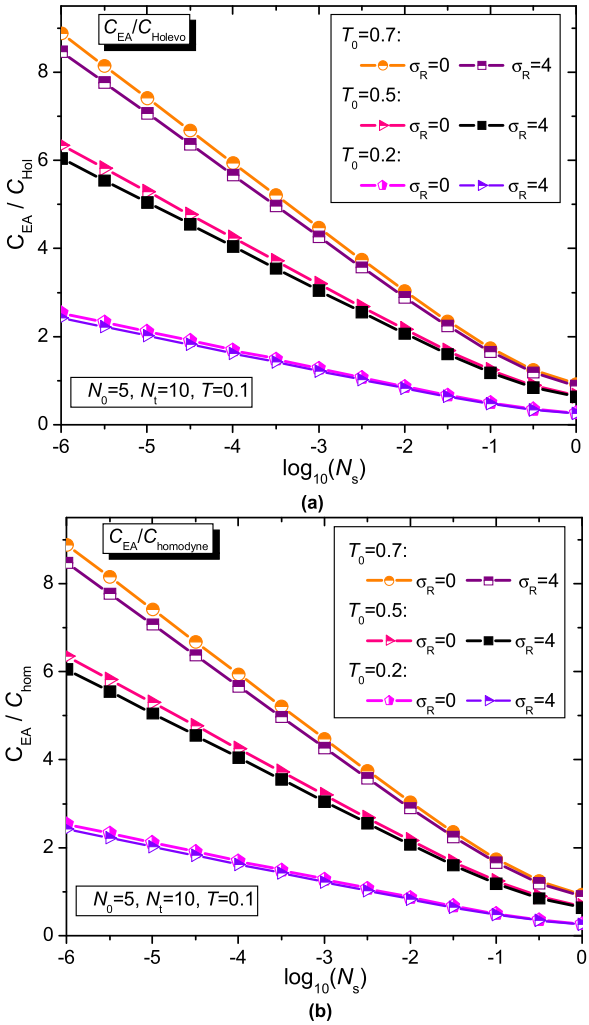


FIGURE 6. The improvement in capacity by entanglement assistance for imperfect pre-shared entanglement distribution in the presence of atmospheric turbulence in main channel: (a) C_{EA}/C_{Hol} and (b) C_{EA}/C_{hom}.

channel is modeled as noisy Bosonic channel with $N_0 = 5$. In addition to turbulence, we assume that the main channel is also affected by diffraction and scattering effects modeled by transmissivity $T = 0.1$. We also assume that in the main channel there exists the background noise with $N_t = 10$. This situation corresponds to Fig. 1(b), where entanglement is distributed over the fiber-based quantum network, and the main channel is an FSO link. Clearly, there is certain degradation in EA capacity compared to the Holevo capacity, in particular when the transmissivity of the entanglement distribution channel is $T_0 \geq 0.5$. On the other hand, in Fig. 6(b) we summarize the EA capacity improvement over homodyne capacity under the same assumptions as in Fig. 6(a). Clearly, the trend is similar as in Fig. 6(a).

To see actual improvement in EA capacity over Holevo and classical capacities for strong turbulence, in Fig. 7 we plot capacities vs. average number of signal photons N_s assuming that entanglement distributed channel is noisy and lossy Bosonic ($N_0 = 2$, $T_0 = 0.5$), while the main channel is affected by strong turbulence ($\sigma_R = 4$), in addition to scattering/diffraction effects ($T = 0.1$) and noise ($N_t = 12$). As long

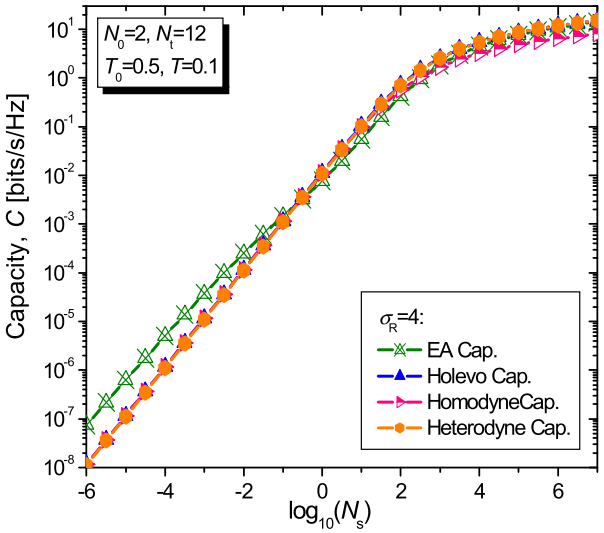


FIGURE 7. **EA capacity vs. classical capacity when main channel is affected by turbulence while entanglement distribution channel is noisy.**

as $N_s < 0.1$, the EA capacity outperforms classical capacities. Holevo and heterodyne capacities are almost identical, while the heterodyne capacity is higher than homodyne capacity for $N_s > 100$.

V. CONCLUDING REMARKS

When the pre-shared entanglement is imperfect, the EA communication in highly noisy environment can be better than classical communication with homodyne detection only if the entanglement distribution channel is less noisy than the main channel and has better transmissivity than the main channel. When both communication and entanglement distribution channels are not used for EA communication but for classical communication instead than the overall classical capacity is always higher than the EA capacity. In the presence of strong atmospheric turbulence in the main channel, the improvement of EA capacity over classical capacity is lower compared to the case without turbulence, but the degradation is not significant.

Even though that the optimum encoding, achieving the EA channel capacity, has been known for decades [4], [5], the design of optimum quantum receiver appears to be still an open problem, although some progress has been made recently [8]. For instance, authors in [8] proposed to use the multiple sections of the feed-forward (FF)-sum-frequency generation (SFG) receiver, initially proposed to detect the target in highly noisy environment [19]. Unfortunately, the complexity of this scheme is too high and has not been experimentally demonstrated yet.

REFERENCES

[1] I. B. Djordjevic, *Physical-Layer Security and Quantum Key Distribution*. Cham, Switzerland: Springer, 2019.

[2] G. Cariolaro, *Quantum Communications*. Cham, Switzerland: Springer, 2015.

[3] I. B. Djordjevic, *Quantum Information Processing, Quantum Computing, and Quantum Error Correction: An Engineering Approach*, 2nd ed. New York, NY, USA: Academic, 2021.

[4] A. Holevo and R. Werner, "Evaluating capacities of bosonic Gaussian channels," *Phys. Rev. A, Gen. Phys.*, vol. 63, no. 3, Feb. 2001, Art. no. 032312.

[5] A. S. Holevo, "On entanglement-assisted classical capacity," *J. Math. Phys.*, vol. 43, no. 9, pp. 4326–4333, 2002.

[6] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem," *IEEE Trans. Inf. Theory*, vol. 48, no. 10, pp. 2637–2655, Oct. 2002.

[7] M. Sohma and O. Hirota, "Capacity of a channel assisted by two-mode squeezed states," *Phys. Rev. A, Gen. Phys.*, vol. 68, no. 2, Aug. 2003, Art. no. 022303.

[8] H. Shi, Z. Zhang, and Q. Zhuang, "Practical route to entanglement-assisted communication over noisy bosonic channels," *Phys. Rev. Appl.*, vol. 13, no. 3, Mar. 2020, Art. no. 034029.

[9] R. Van Meter and S. J. Devitt, "The path to scalable distributed quantum computing," *Computer*, vol. 49, no. 9, pp. 31–42, Sep. 2016.

[10] Z. Zhang and Q. Zhuang, "Distributed quantum sensing," 2020, *arXiv:2010.14744*. [Online]. Available: http://arxiv.org/abs/2010.14744

[11] A. M. Childs, "Secure assisted quantum computation," *Quantum Inf. Comput.*, vol. 5, no. 6, pp. 456–466, 2005.

[12] R. Simon, "Peres–Horodecki separability criterion for continuous variable systems," *Phys. Rev. Lett.*, vol. 84, no. 12, p. 2726, 2000.

[13] L.-M. Duan, G. Giedke, J. I. Cirac, and P. Zoller, "Inseparability criterion for continuous variable systems," *Phys. Rev. Lett.*, vol. 84, no. 12, p. 2722, 2000.

[14] A. Serafini, "Multimode uncertainty relations and separability of continuous variable states," *Phys. Rev. Lett.*, vol. 96, no. 11, Mar. 2006, Art. no. 110402.

[15] S. Pirandola, A. Serafini, and S. Lloyd, "Correlation matrices of two-mode bosonic systems," *Phys. Rev. A, Gen. Phys.*, vol. 79, no. 5, May 2009, Art. no. 052327.

[16] C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd, "Gaussian quantum information," *Rev. Mod. Phys.*, vol. 84, no. 2, pp. 621–669, 2012.

[17] I. B. Djordjevic, *Advanced Optical and Wireless Communications Systems*. Cham, Switzerland: Springer, 2018.

[18] L. C. Andrews and R. L. Philips, *Laser Beam Propagation Through Random Media*. Bellingham, WA, USA: SPIE, 2005.

[19] Q. Zhuang, Z. Zhang, and J. H. Shapiro, "Optimum mixed-state discrimination for noisy entanglement-enhanced sensing," *Phys. Rev. Lett.*, vol. 118, no. 4, Jan. 2017, Art. no. 040801.

[20] V. Nafria, X. Han, and I. B. Djordjevic, "Improving free-space optical communication with adaptive optics for higher order modulation," *Proc. SPIE*, vol. 11509, Aug. 2020, Art. no. 115090K.

[21] V. Nafria, C. Cui, I. B. Djordjevic, and Z. Zhang, "Adaptive-optics enhanced distribution of entangled photons over turbulent free-space optical channels," in *Proc. CLEO*, 2021, May 2021.

IVAN B. DJORDJEVIC (Fellow, IEEE) received the Ph.D. degree from the University of Nis, Yugoslavia, in 1999.

He is currently a Professor of electrical and computer engineering and optical sciences with the University of Arizona; the Director of the Optical Communications Systems Laboratory (OCSL) and Quantum Communications (QuCom) Laboratory; and the Co-Director of the Signal Processing and Coding Laboratory. Prior to joining the University of Arizona, he held appointments with the University of Bristol; the University of the West of England, U.K.; Tyco Telecommunications, USA; the National Technical University of Athens, Greece; and State Telecommunication Company, Yugoslavia. He has authored or coauthored eight books, more than 540 journal and conference publications, and holds 54 U.S. patents.

Dr. Djordjevic is also an OSA Fellow. Since 2016, he has been serving as an Associate Editor/a member of editorial board for *IOP Journal of Optics* and *Physical Communication Journal* (Elsevier). He also serves as an Area Editor/an Associate Editor/a member of editorial board for the journals, such as IEEE Communications Letters, OSA/IEEE Journal of Optical Communications and Networking, *Optical and Quantum Electronics*, and *Frequenz*.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09436-y OPEN

Unifying scrambling, thermalization and entanglement through measurement of fidelity out-of-time-order correlators in the Dicke model

R.J. Lewis-Swan[1,2], A. Safavi-Naini[1,2], J.J. Bollinger[3] & A.M. Rey[1,2]

Scrambling is the process by which information stored in local degrees of freedom spreads over the many-body degrees of freedom of a quantum system, becoming inaccessible to local probes and apparently lost. Scrambling and entanglement can reconcile seemingly unrelated behaviors including thermalization of isolated quantum systems and information loss in black holes. Here, we demonstrate that fidelity out-of-time-order correlators (FOTOCs) can elucidate connections between scrambling, entanglement, ergodicity and quantum chaos (butterfly effect). We compute FOTOCs for the paradigmatic Dicke model, and show they can measure subsystem Rényi entropies and inform about quantum thermalization. Moreover, we illustrate why FOTOCs give access to a simple relation between quantum and classical Lyapunov exponents in a chaotic system without finite-size effects. Our results open a path to experimental use FOTOCs to explore scrambling, bounds on quantum information processing and investigation of black hole analogs in controllable quantum systems.

[1] JILA, NIST and Department of Physics, University of Colorado, Boulder, CO 80309, USA. [2] Center for Theory of Quantum Matter, University of Colorado, Boulder, CO 80309, USA. [3] NIST, Boulder, CO 80305, USA. These authors contributed equally: R. J. Lewis-Swan, A. Safavi-Naini. Correspondence and requests for materials should be addressed to A.M.R. (email: arey@jilau1.colorado.edu)

Recent studies have shown that isolated many-body quantum systems, under unitary time evolution, can become highly entangled and thus thermalize. This understanding has led to insights as to how statistical mechanics emerges in closed quantum systems[1–3]. Moreover, the relevance of entanglement as a resource for quantum information processing, quantum communication and metrology has stimulated cross-disciplinary efforts to quantify and characterize entanglement. Experimental progress in controlling clean, highly isolated, and fully tunable quantum systems, where entanglement can be measured, have resulted in radical advances in this direction. However, such measurements have been restricted to few body systems, including arrays of 6×2 bosonic atoms[4], three superconducting qubits[5], and systems of $\lesssim$20 trapped ions[6,7]. The model we study here and the measurements we propose can be implemented in trapped ions with more than 100 spins.

Concurrently, out-of-time-order correlations (OTOCs)[8–13]

$$F(t) = \langle \hat{W}^\dagger(t)\hat{V}^\dagger\hat{W}(t)\hat{V}\rangle, \quad (1)$$

have been identified as measures of the dynamics of quantum information scrambling. Here, $\hat{W}(t) = e^{i\hat{H}t}\hat{W}e^{-i\hat{H}t}$, with $\hat{H}$ a quantum many-body Hamiltonian, and $\hat{W}$ and $\hat{V}$ two initially commuting and unitary operators. While OTOCs can be computed with respect to any (possibly mixed) state, here we focus on the case where the initial state of the system is pure. The quantity $\mathrm{Re}[F(t)] = 1 - \langle[\hat{V}^\dagger, \hat{W}^\dagger(t)][\hat{W}(t), \hat{V}]\rangle/2$ encapsulates the degree that $\hat{W}(t)$ and $\hat{V}$ fail to commute at later times due to the time evolution of $\hat{W}$ under $\hat{H}$. (If $\hat{V}$ is not unitary but a projector, e.g. $\hat{V}\hat{V}^\dagger = \hat{V}$ and $\hat{V}$ commutes with the density matrix of the initial state, then $\mathrm{Re}[F(t)] = 1 - \langle[\hat{V}^\dagger, \hat{W}^\dagger(t)][\hat{W}(t), \hat{V}]\rangle$.) The fastest scramblers[8–10,14], such as black holes, feature an exponential growth of scrambling which manifests as $1 - \mathrm{Re}[F(t)] \sim e^{\lambda_Q t}$. Here, λ_Q is the quantum Lyapunov exponent that serves as a proxy for quantum chaos. Regardless of the OTOCs' apparent complexity[13,15–17], the capability to perform many-body echoes (see Fig. 1) in current experiments[18–21] has opened a path for the experimental investigation of quantum scrambling; however, so far those have not probed quantum chaos or fast scrambling.

Here we show that fidelity out-of-time-order correlators (FOTOCs), a specific family of fidelity out-of-time-order correlators, which set $\hat{V}$ to be a projector on the initial state, can provide profound insight on scrambling behavior. We explicitly compute FOTOCs in the in the Dicke model[22], an iconic model in quantum optics, and illustrate how FOTOCs elucidate theoretical connections between scrambling, volume-law Rényi entropy (RE) and thermalization, while linking quantum and classical chaos (Fig. 1a). Additionally, we discuss how one can probe these connections readily in experiments.

Results

Model. The Dicke model (DM)[22] describes the coupling of a single large spin and a harmonic oscillator and has been recently implemented in atomic[23–26] and trapped ion setups[27]. The Hamiltonian of the DM is given by

$$\hat{H}_D = \frac{2g}{\sqrt{N}}\left(\hat{a} + \hat{a}^\dagger\right)\hat{S}_z + \delta\hat{a}^\dagger\hat{a} + B\hat{S}_x, \quad (2)$$

where B characterizes the strength of the transverse field, δ the detuning of the bosonic mode from the driving field with strength g that generates the spin−boson coupling. Here, g, δ, $B \geq 0$. The operator $\hat{a}$ ($\hat{a}^\dagger$) is the bosonic annihilation (creation) operator of the mode, and $\hat{S}_\alpha = \sum_{j=1}^{N}\hat{\sigma}_j^\alpha 2$ are collective spin operators with $\hat{\sigma}_j^\alpha$ ($\alpha = x, y, z$) the Pauli matrices for the jth spin-1/2.

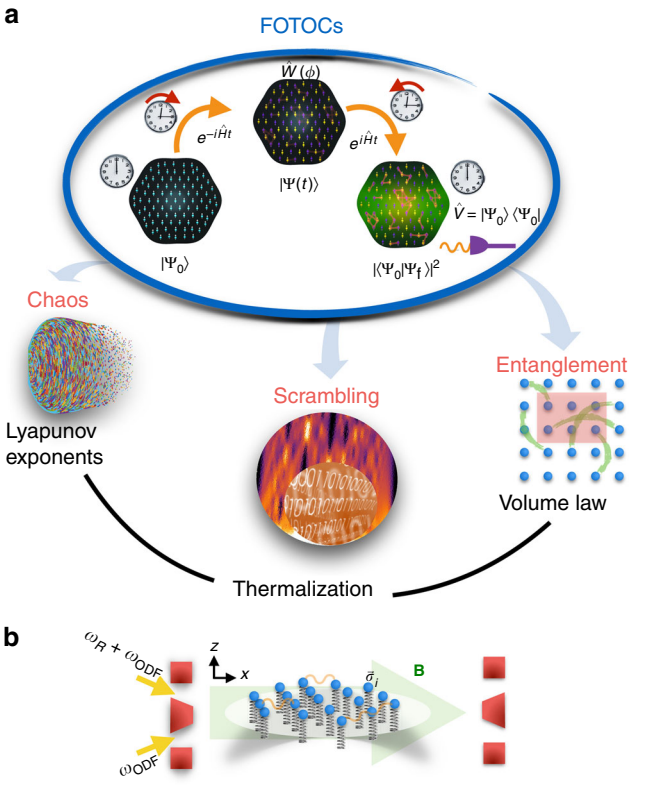


Fig. 1 Unifying chaos, scrambling, entanglement and thermalization through the measurement of fidelity out-of-time-order correlators (FOTOCs). **a** Scheme: an initial state, $|\psi_0\rangle$ is evolved under an interacting Hamiltonian $\hat{H}$ for a time t. Inverting the sign of $\hat{H}$ and evolving again for time t to the final state $|\psi_f\rangle$, implements the many-body time-reversal, which ideally takes the system back to the initial state $|\psi_0\rangle$. If a perturbation $\hat{W}(\phi)$ is inserted between the two halves of the time evolution and the many-body overlap with the initial state is measured at the end of the protocol, $\hat{V} = |\Psi_0\rangle\langle\Psi_0|$, then a special type of fidelity OTOC (FOTOC) is implemented. **b** The Dicke model is engineered in a Penning trap ion crystal by applying a pair of lasers, resonant only with the center-of-mass mode, to generate the spin−phonon interaction and resonant microwaves to generate the transverse field

Connections between scrambling dynamics and chaos. Even when restricted to the Dicke manifold, i.e. states with $S = N/2$, with $S(S+1)$ the eigenvalue of the total spin operator $\hat{S}^2 = \hat{S}_x^2 + \hat{S}_y^2 + \hat{S}_z^2$, this model exhibits rich physics (see Fig. 2a). At zero temperature, $T = 0$, the DM features a quantum phase transition (QPT) as the system crosses a critical field $B_c = 4g^2/\delta$. For $B > B_c$ (normal phase), the ground-state is described by spins aligned along the transverse field and a bosonic vacuum. For $B < B_c$ (superradiant phase), the ground-state is ferromagnetic, $\langle|\hat{S}_Z|\rangle \sim N/2$, and characterized by macroscopic occupation of the bosonic mode (Fig. 2a). Furthermore, in the superradiant phase ($B < B_c$), the DM features a family of excited-state quantum phase transitions (ESQPTs). The ESQPTs are signaled by singularities in the energy-level structure and a change in the spectral statistics[28–31] at a critical energy $E_c = -BN/2$ that coincides with the ground-state energy of the normal phase. Figure 2a shows how the nearest-neighbor spacing distribution $P(s)$, where s is a normalized distance between two neighboring energy levels, features a different character on either side of E_c. For $E > E_c$ the spectral statistics are similar to the Wigner−Dyson distribution $P_W(s) = \pi s/2\exp(-\pi s^2/4)$, which in random-matrix theory describes a chaotic system. For $E < E_c$ the shape of the histograms

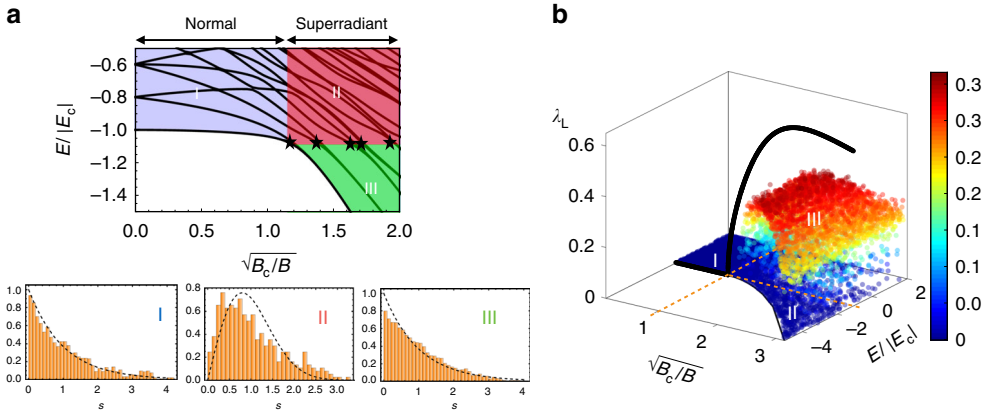


Fig. 2 Characterization of classical and quantum chaos in the Dicke model. **a** Phase diagram of the Dicke model. At zero temperature it exhibits a quantum phase transition between a normal to a superradiant phase, at $B = B_c$. A line of excited energy quantum phase transitions (ESQPTs) occurs at the critical energy $E_c = -BN/2$, signaled by singularities in the energy level structure (indicated by stars). Note that for figure clarity we have used a small system $N = 20$ resulting in the small deviation of the ESQPTs from $E_c = -BN/2$. The ESQPTs are accompanied by a change in the level statistics which we denote by (I) −(III) (note that no eigenstates exist in the unlabeled white region). For (II) and (III) the spectrum is divided into low and high energy parts, separated by the ESQPT at $E = E_c$, from which the statistics $P(s)$, where s is the level spacing, are computed separately. (I) exhibits Poissonian statistics (regular regime), while (II) displays statistics similar to a Wigner−Dyson distribution indicative of level repulsion and quantum chaos, and (III) exhibits a mixture of both. The numerical parameters are $g/(2\pi) = 0.66$ kHz and $\delta/(2\pi) = 0.5$ kHz. **b** Lyapunov exponents for the mean-field dynamics of an ensemble of random states sorted by normalized mean-field energy $E/|E_c|$ with $E_c = -BN/2$, as a function of the field $\sqrt{B_c/B}$ relative to critical field $B_c = 4g^2/\delta$. A crossover between regular ($B > B_c$) and chaotic dynamics ($B < B_c$) characterized by $\lambda \simeq 0$ and $\lambda_L > 0$ respectively, occurs at $B = B_c$. For $B > B_c$ and energies $E \lesssim E_c$ the dynamics becomes increasingly regular. Source data are provided as a Source Data file

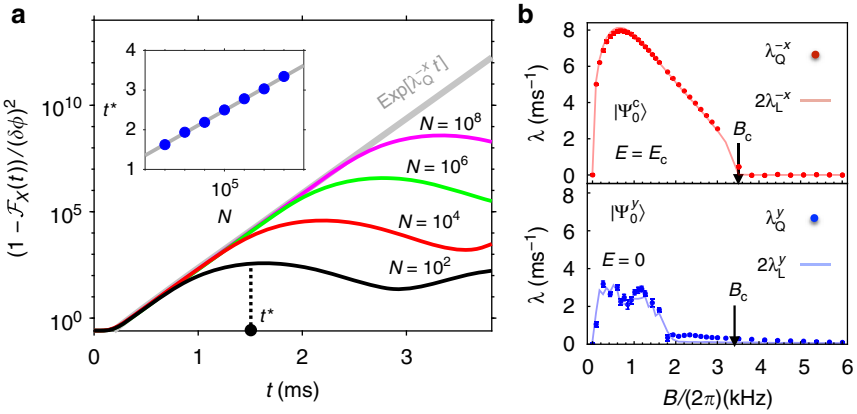


Fig. 3 Signatures of classical chaos in quantum FOTOCs. **a** Initial exponential growth of the FOTOC, $[1 - \mathcal{F}_X(t)]/(\delta\phi)^2$ and the initial state $|\Psi_0^c\rangle = |(-N/2)_x\rangle \otimes |0\rangle$ (see Supplementary Note 1 for examples of exponential growth in other states). We assume $\delta\phi \ll 1/N$ such that we may equivalently use $\mathrm{var}(\hat{X}) \simeq [1 - \mathcal{F}_X(t)]/(\delta\phi)^2$ for the plotted data. The scrambling time t^* is defined by the saturation of the FOTOC, which we extract from the first maximum and plot in the inset (blue data). We find $t^* \sim a_0 + \log(N)/\lambda_Q$ with a_0 a fit parameter (gray line). **b** Lyapunov exponent, λ, as a function of transverse field: Quantum λ_Q (red markers) and classical $2\lambda_L$ (solid lines). Superscript notation of the exponents denotes the initial polarization of the chosen coherent spin state. Top panel for $|\Psi_0^c\rangle$, the same state as (**a**), and bottom for $|\Psi_0^y\rangle \equiv |(-N/2)_y\rangle \otimes |0\rangle$, here $N = 10^4$ particles. In both plots we observe $\lambda_Q \simeq 2\lambda_L$. Error bars for λ_Q are a 95% confidence interval from an exponential fitted to the numerical data. Coupling g and detuning δ are same as Fig. 2. In (**a**) $B/(2\pi) = 0.7$ kHz ($B/B_c = 0.2$). Source data are provided as a Source Data file

is neither Wigner−Dyson nor Poissonian $P_P(s) = \exp(-s)$. The latter characterizes level statistics of non-ergodic systems, and is observed in the normal phase. While the deviations from clear Wigner−Dyson or Poissonian statistics in regimes II and III are attributable to finite-size effects[28], we emphasize that even for this small system they clearly show a stark contrast in the degree of level repulsion, which is a qualitative signature of quantum chaos.

Similar features appear in the classical dynamics of the DM[30,32–35], manifested in the different behavior of trajectories in phase-space computed from the mean-field equations of motion for: $\tilde{\mathbf{x}} = (\langle\hat{S}_x\rangle, \langle\hat{S}_y\rangle, \langle\hat{S}_z\rangle, \alpha_R, \alpha_I)$, where $\langle \dots \rangle$ denotes the expectation values, and $\alpha_{R(I)}$ is the real (imaginary) part of $\langle\hat{a}\rangle$. In the superradiant phase and for mean-field energies $E > E_c$, two trajectories initially separated by $\Delta\tilde{\mathbf{x}}(0)$ in phase-space diverge as $|\Delta\tilde{\mathbf{x}}(t)| \sim |\Delta\tilde{\mathbf{x}}(0)|e^{\lambda_L t}$ at sufficiently long times[36]. The exponential growth, associated with a positive Lyapunov exponent $\lambda_L > 0$, diagnoses chaos in a classical system. In Fig. 2b we show the maximal Lyapunov exponent for an ensemble of random initial product states as a function of the transverse field and the normalized mean-field energy E/E_c (see Methods). For $E < E_c$ in the superradiant phase ($B < B_c$) and all energies in the normal phase ($B > B_c$), the Lyapunov exponent is small or zero, consistent with the Poissonian character of the quantum-level statistics in this parameter regime[34,35]. For $E > E_c$ and $B < B_c$ a positive exponent is found signaling chaos. Note that the state $|\Psi_0^c\rangle = |(-N/2)_x\rangle \otimes |0\rangle$, where $\hat{S}_x|(-N/2)_x\rangle = (-N/2)|(-N/2)_x\rangle$, lies exactly at the ESQPT

critical energy, $\langle\Psi_0^c|\hat{H}_D|\Psi_0^c\rangle = E_c$, and possesses the largest classical λ_L (see Fig. 2).

In quantum systems OTOCs may serve as a diagnostic for quantum chaos. However, such diagnosis has proved difficult, since any exact numerical treatment is only possible in small systems, where many-body observables saturate quickly at the Ehrenfest time given by $\lambda_Q t^* \sim \log N$, at which the quantum information is thoroughly lost to a "local" observer. Here we demonstrate that we can overcome this limitation and compute OTOCs for macroscopic systems if, for a Hermitian operator $\hat{G}$, one restricts $\hat{W}_G = e^{i\delta\phi\hat{G}}$ to be a sufficiently small perturbation ($\delta\phi \ll 1$) and sets $\hat{V}$ to be a projection operator onto a simple initial state $|\psi_0\rangle$, i.e. $\hat{V} = \hat{\rho}(0) = |\Psi_0\rangle\langle\Psi_0|$. This is because in the perturbative limit $\delta\phi \ll 1$, this particular type of fidelity OTOC (FOTOC)[18,19], $\mathcal{F}_G(t,\delta\phi) \equiv \langle \hat{W}_G^\dagger(t)\hat{\rho}(0)\hat{W}_G(t)\hat{\rho}(0)\rangle$ (such that for a pure state $\mathcal{F}_G(t) \equiv |\langle\psi_0|e^{i\hat{H}t}e^{i\delta\phi\hat{G}}e^{-i\hat{H}t}|\psi_0\rangle|^2$) reduces to[37]

$$1 - \mathcal{F}_G(t,\delta\phi) \approx \delta\phi^2(\langle\hat{G}^2(t)\rangle - \langle\hat{G}(t)\rangle^2) \equiv \delta\phi^2 \mathrm{var}[\hat{G}(t)], \quad (3)$$

where $\mathrm{var}[\hat{G}(t)]$ is the variance of $\hat{G}$. This relation establishes a connection between the exponential growth of quantum variances and quantum chaos, enables us to visualize the scrambling dynamics of a quantum system using a semi-classical picture[38] and to map the FOTOC to a two-point correlator which can be computed using well-known phase-space methods, such as the truncated Wigner approximation (see Methods)[39,40]. We observe perfect agreement between the exact dynamics of the FOTOC with the associated variance, $\mathrm{var}(\hat{G})$ for sufficient small $\delta\phi$, enabling us to use phase-space methods to compute the FOTOCs in a parameter regime inaccessible to exact numerical diagonalization where exponential scrambling can be clearly identified.

Moreover, it provides a link between the FOTOCs and the quantum Fisher information (QFI)[19,41–43], as the variance of $\hat{G}$ is proportional to the QFI of a pure state, whilst for a mixed state the variance gives a lower bound on the QFI. Note that in the latter case FOTOCs are defined by replacing $\hat{V}$ by the initial density matrix $|\Psi_0\rangle\langle\Psi_0| \to \hat{\rho}_0$, and expectation values are computed by appropriate traces. The QFI quantifies the maximal precision with which a parameter $\delta\phi$ in the unitary of $\hat{W}$ can be estimated using an interferometric protocol with an input quantum state $|\psi(t)\rangle$, while simultaneously serving as a witness to multipartite entanglement[19,44–46].

In Fig. 3a we plot the FOTOCs of a small perturbation using $\hat{G} = \hat{X} = \frac{1}{2}(\hat{a} + \hat{a}^\dagger)$ starting with $|\Psi_0\rangle = |\Psi_0^c\rangle$. In the superradiant phase we observe that after a short time of slow dynamics, $t_\lambda \sim \lambda_Q^{-1}$, the FOTOCs feature an exponential growth $\sim e^{\lambda_Q t}$, before saturating at $t^* \sim \log N$ (see inset). The quantum exponent is found to be independent of system size N. For this initial state, and all the product states we have investigated numerically (Supplementary Note 1), we have observed that $\lambda_Q \simeq 2\lambda_L$, as shown in Fig. 3b. Indeed, for any $\hat{G}$ that corresponds to a linear function of the classical phase-space variables (see Methods and Supplementary Note 1), the quantum exponent should be related to the classical Lyapunov exponent by this relation. A similar factor of two relating the classical and quantum exponents has previously been observed in refs. [37,47]. This correspondence can be explained by semi-classical arguments (see Methods), and the numeric prefactor is attributable to the definition of the classical Lyapunov exponent in terms of a distance in phase-space, while the FOTOC reduces to the quantum variance.

FOTOCs as a probe of entanglement and quantum thermalization. We now move beyond the semi-classical arena and explore connections between FOTOCs and entanglement entropy. In a closed system S the second-order Rényi entropy, RE, $S_2(\hat{\rho}_A) = -\log\mathrm{Tr}(\hat{\rho}_A^2)$ measures the entanglement between a subsystem A and its complement $A_c = S - A$, with $\hat{\rho}_A$ the reduced density matrix of A after tracing over A_c. Although scrambling and entanglement buildup are closely connected, they are not the same. Nevertheless, a formal relationship between the OTOCs and $S_2(\hat{\rho}_A)$ exists[11], which requires averaging OTOCs over a complete basis of operators of the system subsystem A. Based on this relation, measuring RE via OTOCs appears as challenging as directly measuring $S_2(\hat{\rho}_A)$. However, this is not always the case. We will show that for collective Hamiltonians, such as the DM, there is a simple correspondence between the Fourier spectrum of FOTOCs and the RE, which facilitates experimental access to $S_2(\hat{\rho}_A)$ via global measurements and collective rotations.

To illustrate the connection we first write the density matrix of the full system in a basis spanned by the eigenstates of the spin operator $\hat{S}_\mathbf{r} \equiv (e_\mathbf{r} \cdot S)$, where $\mathbf{e_r}$ is a unit vector in the Bloch sphere, satisfying $\hat{S}_\mathbf{r}|m_\mathbf{r}\rangle = m_\mathbf{r}|m_\mathbf{r}\rangle$, and $\hat{n}|n\rangle = |n\rangle$ the mode number operator $\hat{n} = \hat{a}^\dagger\hat{a}$, i.e. $\hat{\rho} = \sum_{\substack{n,n' \\ m_\mathbf{r},m_\mathbf{r}'}} \rho^{n',n}_{m_\mathbf{r}',m_\mathbf{r}} |n'\rangle\langle n| \otimes |m_\mathbf{r}'\rangle\langle m_\mathbf{r}|$. We adopt a convention for the coefficients of the density matrix elements where superscripts are associated with the bosonic mode, and subscripts with the spin. In this basis the density matrix can be divided into blocks, $\hat{\rho} = \sum_M \hat{\rho}_M^{\hat{S}_\mathbf{r}}$ with $\hat{\rho}_M^{\hat{S}_\mathbf{r}} = \sum_{\substack{n,n' \\ m_\mathbf{r}}} \rho^{n',n}_{m_\mathbf{r}+M,m_\mathbf{r}} |n'\rangle\langle n| \otimes |m_\mathbf{r} + M\rangle\langle m_\mathbf{r}|$, in such a way that $\hat{\rho}_M^{\hat{S}_\mathbf{r}}$ contains all coherences between states with spin eigenvalues that differ by M. A similar decomposition can be performed in terms of the bosonic coherences as $\hat{\rho} = \sum_M \hat{\rho}_M^{\hat{n}}$ with $\hat{\rho}_M^{\hat{n}} = \sum_{n m_\mathbf{r},m_\mathbf{r}'} \rho^{n+M,n}_{m_\mathbf{r}',m_\mathbf{r}} |n+M\rangle\langle n| \otimes |m_\mathbf{r}'\rangle\langle m_\mathbf{r}|$. Associated with this representation one can define the so-called multiple quantum intensities $I_M^{\hat{G}} = \mathrm{Tr}[\hat{\rho}_{-M}^{\hat{G}}\hat{\rho}_M^{\hat{G}}]$. Of particular interest for us are the $I_0^{\hat{G}}$ components which are "incoherent" with respect to $\hat{G}$.

The intensities $I_M^{\hat{G}}(t)$ can be accessed experimentally from FOTOCs via the relation $\mathcal{F}_G(t,\phi) = \sum_M I_M^{\hat{G}}(t)e^{-iM\phi}$[18,19,48–50] by choosing $\hat{W}_G(\phi) = e^{-i\phi\hat{G}}$ and $\hat{G} = \hat{S}_\mathbf{r}$ or $\hat{G} = \hat{n}$, i.e. collective spin or boson rotations respectively. In terms of the $I_M^{\hat{G}}(t)$ the entanglement between the spins and the phonons characterized by the purity $\mathrm{Tr}\left[\hat{\rho}_{\mathrm{ph}}^2\right] = \sum_{\substack{n,n' \\ m_\mathbf{r},m_\mathbf{r}'}} \rho^{n,n'}_{m_\mathbf{r},m_\mathbf{r}} \rho^{n',n}_{m_\mathbf{r}',m_\mathbf{r}'}$ can be written as

$$\mathrm{Tr}\left[\hat{\rho}_{\mathrm{ph}}^2(t)\right] \equiv I_0^{\hat{S}_\mathbf{r}}(t) + I_0^{\hat{n}}(t) - D_{\mathrm{diag}}^{\hat{S}_\mathbf{r},\hat{n}}(t) + C_{\mathrm{off}}^{\hat{S}_\mathbf{r},\hat{n}}(t). \quad (4)$$

The terms $D_{\mathrm{diag}}^{\hat{S}_\mathbf{r},\hat{n}}(t)$ and $C_{\mathrm{off}}^{\hat{S}_\mathbf{r},\hat{n}}(t)$ are explicitly detailed in the Methods, but importantly $D_{\mathrm{diag}}^{\hat{S}_\mathbf{r},\hat{n},}(t)$ is composed purely of the diagonal elements of $\hat{\rho}$ while $C_{\mathrm{off}}^{\hat{S}_\mathbf{r},\hat{n}}(t)$ contains information about coherences. During unitary evolution the characteristic dephasing time of the coherences is $t_c \sim \lambda_Q^{-1}$, which for scrambling systems is much faster than $t^* \sim \lambda_Q^{-1}\log N$. After t_c any remaining coherences are fully randomized and destructively interfere yielding $C_{\mathrm{off}}^{\hat{S}_\mathbf{r},\hat{n}} \to 0$. This feature, together with the fact that for those systems also the magnitude of $D_{\mathrm{diag}}^{\hat{S}_\mathbf{r},\hat{n}}$ becomes much smaller than $I_0^{\hat{S}_\mathbf{r}}$ and $I_0^{\hat{n}}$ as the density matrix spreads out over the systems degrees of freedom, allows us to approximate $\mathrm{Tr}\left[\hat{\rho}_{\mathrm{ph}}^2(t)\right] \approx I_0^{\hat{S}_\mathbf{r}}(t) + I_0^{\hat{n}}(t)$. While at $t < t_c$ these conditions are not necessarily satisfied, we still find that there can be a correspondence between the FOTOCs and RE by picking a state that is fully incoherent at time $t = 0$, $C_{\mathrm{off}}^{\hat{S}_\mathbf{r},\hat{n}}(0) = 0$. An example of

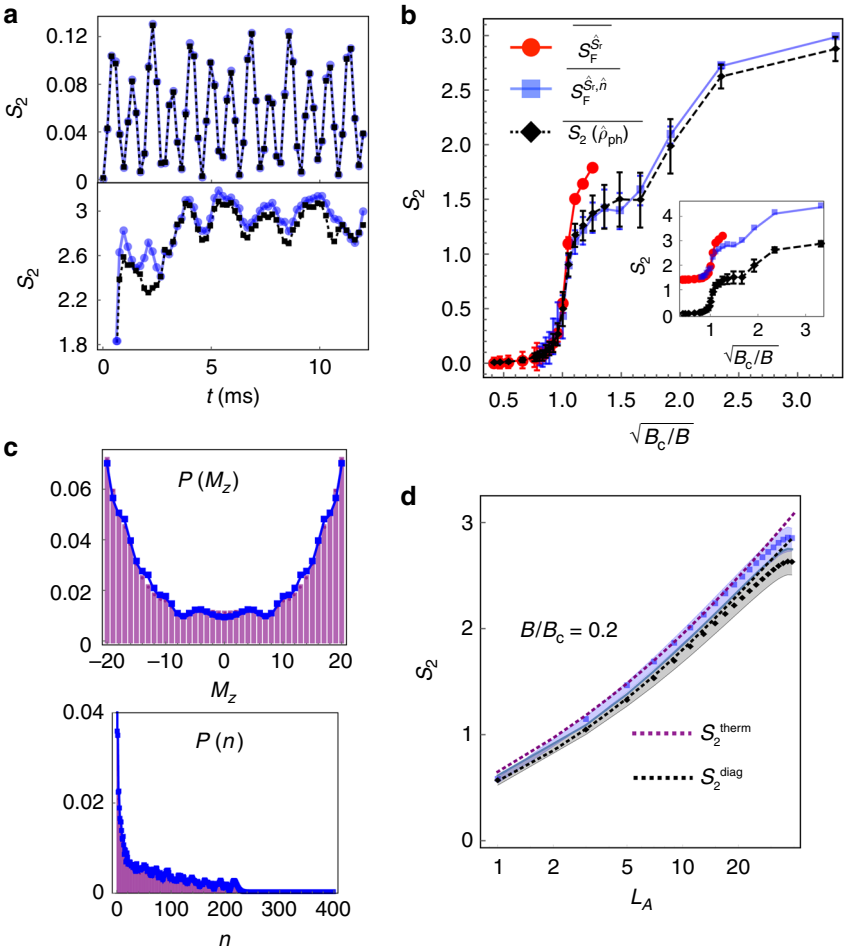


Fig. 4 Using RE and FOTOCs to characterize chaos and thermalization in the Dicke model. **a** Time evolution of the spin−phonon RE $S_2(\hat{\rho}_{ph})$ (black lines) for the initial state $|\Psi_0^c\rangle = |(-N/2)_x\rangle \otimes |0\rangle$ with $B > B_c$ (top) and $B < B_c$ (bottom). The RE is tracked excellently by the FOTOC expressions (blue lines) $S_F^{\hat{S}_r} = -\log(I_0^{\hat{S}_r})$ and $S_F^{\hat{S}_r,\hat{n}} = -\log(I_0^{\hat{S}_r} + I_0^{\hat{n}})$ respectively. Here, $\hat{S}_r$ is chosen to minimize the coherence and diagonal terms in Eq. (4) (Supplementary Methods). **b** Long-time spin−phonon RE $S_2(\hat{\rho}_{ph})$ as a function of transverse field. To remove finite-size effects and residual oscillations we plot a time-averaged value $\overline{S_2(\hat{\rho}_{ph})}$ for 4 ms $\leq t \leq$ 12 ms (FOTOC quantities are averaged identically). The regular and chaotic dynamics for the initial state $|\Psi_0^c\rangle$ are clearly delineated: $\overline{S_2(\hat{\rho}_{ph})} \approx 0$ for $B > B_c$ and $\overline{S_2(\hat{\rho}_{ph})} > 0$ for $B < B_c$ respectively. Error bars indicate standard deviation of temporal fluctuations. In the inset we plot the same FOTOC quantities but including decoherence due to single-particle dephasing at the rate $\Gamma = 60\ s^{-1}$. The coherent parameters g, B and δ are enhanced by a factor of 16 compared to the main panel, as per ref. [56]. **c** Time-averaged distribution functions (markers) for spin-projection $P(M_z)$ and phonon occupation $P(n)$ (6 ms $\leq t \leq$ 12 ms). We compare to the distribution of the diagonal ensemble (purple bars, see Methods). **d** Bipartite RE $S_2(\hat{\rho}_{L_A})$ (black markers) as a function of partition size L_A of the spins, averaged over same time window as (**c**). For comparison, we plot the RE of a thermal canonical ensemble with corresponding temperature T fixed by the energy of the initial state $|\Psi_0^c\rangle$, S_2^{therm} and the RE of the diagonal ensemble (see Methods). Volume-law behavior of the RE is replicated by the FOTOC quantity (blue markers). Note that the dimension of the spin Hilbert space scales linearly with L_A. Shaded regions indicate standard deviation of temporal fluctuations. Data for (**a**)-(**d**) is obtained for $N = 40$, with g and δ identical to calculations of Fig. 2. For (**c**) and (**d**) we choose $B/(2\pi) = 0.7$ kHz ($B/B_c = 0.2$). Source data are provided as a Source Data file

such a state is $|\Psi_0^c\rangle$ and $\hat{G} = \hat{S}_x$. This choice enforces the $C_{off}^{\hat{S}_r,\hat{n}}$ term to remain small at short times. Moreover, for $|\Psi_0\rangle$ we find it is also possible to access $S_2(\hat{\rho}_{ph})$ via $I_0^{\hat{S}_r}$ even in the regime $B > B_c$, where no scrambling occurs. This is because the contributions from $I_0^{\hat{n}}$ and $D_{diag}^{\hat{S}_r,\hat{n}}$ cancel and $\mathrm{Tr}\left[\hat{\rho}_{ph}^2(t)\right] \approx I_0^{\hat{S}_r}$.

In Fig. 4a we show the typical behavior of the RE, $S_2(\hat{\rho}_{ph})$, in the two different phases for $|\Psi_0\rangle$. First, in the normal phase (panel (i)), $B \ll B_c$, the dynamics is dominated by precession about the transverse field and the entanglement entropy exhibits small amplitude oscillations[51]. Conversely, in the superradiant phase (panel (ii)) $B \ll B_c$ we observe a rapid growth of entanglement and saturation past the transient regime. We summarize our results in Fig. 4b where we plot the time-averaged value of $S_2\left(\hat{\rho}_{ph}\right)$ vs. $\sqrt{B_c/B}$. We associate the fast growth of $S_2\left(\hat{\rho}_{ph}\right)$ at $B/B_c \sim 1$ with a crossover from the integrable to the chaotic regime. To further illustrate this connection, we compare the approximate RE obtained via $S_F^{\hat{S}_r,\hat{n}} \equiv -\log[I_0^{\hat{S}_r}(t) + I_0^{\hat{n}}(t)]$ and $S_F^{\hat{S}_r} \equiv -\log[I_0^{\hat{S}_r}]$ with the exact RE in Fig. 4b. It is observed that in all parameter regimes one can make a quantitative link between the RE and FOTOCs, especially under proper optimization of the rotation axis $\hat{S}_r$ at each time to minimize the coherence and diagonal terms in Eq. (4) (see Methods and Supplementary Methods).

The saturation of $S_2(\hat{\rho}_{ph})$ for $B < B_c$ is a signature of thermalization. One can test how "thermalized" the quantum system is by comparing the behavior of the spin and phonon distributions in the long time limit with those of the corresponding diagonal ensemble, characterized by a mixed density matrix $\hat{\rho}_D$ with purely diagonal elements (see Methods)[1–3]. These

comparisons are shown in Fig. 4c, where the time evolved distributions and the ones drawn from the diagonal ensemble are almost indistinguishable.

We can also investigate the growth of entanglement on different size bipartitions for $B < B_c$. For that we split the spin system into a subsystem of size $L_A \leq N$ and evaluate $S_2(\hat{\rho}_{L_A})$ by computing the reduced density matrix $\hat{\rho}_{L_A}$ by tracing over the bosonic degree of freedom and the remaining $N - L_A$ spins. To demonstrate the entanglement grows with system size in a manner consistent with an equivalent thermal state we plot the predictions of a canonical ensemble (see Methods). We observe volume-law entanglement growth for $L_A \ll N$ (see Fig. 4d). However, for $L_A \sim N$ the entanglement growth deviates from this simple prediction. These deviations occur as the full state of the system is pure, and thus eventually one needs to recover $S_2(\hat{\rho}) = 0$, requiring a negative curvature. To demonstrate the intertwined nature of thermalization and the buildup of entanglement we plot the predictions of a canonical ensemble indicated by the dotted purple line (see Methods). We note that FOTOCs can also be used to probe this scaling of the RE with subsystem size. To this end, both $\hat{V}$ and $\hat{W}$ should be restricted to a partition of size L_A of the system, but otherwise the corresponding multiple quantum intensities are computed as discussed above (see also Methods). Figure 4d shows the excellent agreement between the partial system FOTOCs (blue squares) and RE (black diamonds), comparisons that illustrate the utility of FOTOCs to characterize complex many-body entanglement.

Experimental implementation in trapped ion simulators. Trapped ions present a promising experimental platform for the investigation of the physics discussed here[27,52,53]. Here we focus on two-dimensional arrays in a Penning trap where a tunable coupling between the ion's spin, encoded in two hyperfine states, and the phononic center-of-mass (COM) mode of the crystal can be implemented by a pair of lasers with a beatnote frequency detuned by δ from the COM mode and far from resonance to all other modes, which remain unexcited (Fig. 1b). In the presence of microwaves (which generate the transverse field) resonant with the spin level splitting, the effective Hamiltonian is of the form of Eq. 2 as benchmarked in refs. [27,54]. The dynamical control of the transverse field and sign of the detuning from the COM mode enables straightforward implementation of a time-reversal protocol to measure FOTOCs[19] (see Fig. 1). Additionally, the many-body echo requires the application of a spin echo π pulse along $e_r = \hat{y}$ which reverses the signs of $\hat{S}_x$ and $\hat{S}_z$ simultaneously.

Our proposal requires the ability for measuring the fidelity of the full spin−phonon state, which we have not yet demonstrated experimentally. However, this will be possible through a generalization of the protocol discussed in ref. [55] (see Methods). Additionally, our proposal can be adversely affected by decoherence present in the experiment. However, the impact of decoherence will be minimized in future experiments by increasing the magnitude of relevant couplings of the DM via parametric amplification of the ions' motion[27,56], thus reducing the ratio of dissipative to coherent evolution. We illustrate the predicted effect of decoherence, which is dominated by single-particle dephasing due to light scattering from the lasers, in the inset of Fig. 4b. We include the enhancement of the coherent parameters via the protocol described in ref. [56] while using the typical experimental decoherence rate of $\Gamma = 60\,\mathrm{s}^{-1}$. The single-particle decoherence is modeled by an exponential decay of the FOTOC components $I_0^{\hat{G}} \to I_0^{\hat{G}} e^{-\Gamma N t}$ (see Methods). The numerical calculation indicates that even with decoherence the crossover between the two regimes at $B \sim B_c$ is still well captured. Due to numerical complexity of solving a master equation we restrict our simulations to $N = 40$ ions.

Discussion

We have demonstrated that FOTOCs connect the fundamental concepts of scrambling, chaos, quantum thermalization, and multipartite entanglement in the DM. While the concepts presented here have been limited to collective Hamiltonians, we believe they can be generalized to more complex many-body models (Supplementary Note 2). For example, FOTOCs could provide an alternative approach for performing efficient measurements of RE in a way comparable to other state-of-the-art methods which have been used to probe entanglement in systems with up to 20 ions[7]. Generically, FOTOCs could serve as an experimental tool capable of uncovering bounds on information transport and computational complexity, and shed light on how classical behaviors in macroscopic systems emerge from purely microscopic quantum effects.

Note added: Upon completion of this manuscript we became aware of the recent preprints[57,58], which present the numerical and analytic investigation of OTOCs in the Dicke model.

Methods

Classical dynamics and equations of motion. The results presented for the classical model in Fig. 3 are obtained from the Heisenberg equations of motion for the operators via a mean-field ansatz, wherein the operators are replaced by the c-number expectation values, i.e., $\hat{S}_j \to \langle \hat{S}_j \rangle$ for $j = x, y, z$ and $\hat{a} \to \langle \hat{a} \rangle$ (where we adopt $\alpha_{\mathrm{R(I)}}$ as the real (imaginary) component of $\langle \hat{a} \rangle$). We thus obtain an equation of motion for $\tilde{\mathbf{x}} = (\langle \hat{S}_x \rangle, \langle \hat{S}_y \rangle, \langle \hat{S}_z \rangle, \alpha_{\mathrm{R}}, \alpha_{\mathrm{I}})$,

$$\frac{\mathrm{d}\tilde{\mathbf{x}}}{\mathrm{d}t} = F(\tilde{\mathbf{x}}), \tag{5}$$

where

$$F(\tilde{\mathbf{x}}) = \begin{pmatrix} -\delta\alpha_{\mathrm{I}} \\ \delta\alpha_{\mathrm{R}} - \frac{2g}{\sqrt{N}}\langle \hat{S}_z \rangle \\ -\frac{4g}{\sqrt{N}}\alpha_{\mathrm{R}}\langle \hat{S}_y \rangle \\ -B\langle \hat{S}_z \rangle + \frac{4g}{\sqrt{N}}\alpha_{\mathrm{R}}\langle \hat{S}_x \rangle \\ B\langle \hat{S}_y \rangle \end{pmatrix}. \tag{6}$$

Lyapunov exponent. The existence of classical chaos can be characterized by the Lyapunov exponent λ_{L}. By definition, classical chaos implies that two initially close trajectories separated by a distance in phase-space $\Delta\tilde{\mathbf{x}}(0) = |\tilde{\mathbf{x}}_1(0) - \tilde{\mathbf{x}}_2(0)|$ diverge exponentially, $|\Delta\tilde{\mathbf{x}}(t)| \approx |\Delta\tilde{\mathbf{x}}(0)| e^{\lambda_{\mathrm{L}} t}$, and thus $\lambda_{\mathrm{L}} > 0$ is a signature of chaotic dynamics.

Formally, the Lyapunov exponent is then defined by taking the limit[36]

$$\lambda_{\mathrm{L}} \equiv \lim_{t\to\infty} \lim_{|\Delta\tilde{\mathbf{x}}(0)|\to 0} \frac{1}{t} \log \frac{|\Delta\tilde{\mathbf{x}}(t)|}{|\Delta\tilde{\mathbf{x}}(0)|}. \tag{7}$$

As the phase-space of our co-ordinate system is bounded, we evaluate Eq. (7) using the tangent-space method[35,59]. Essentially, rather than monitoring the physical separation $|\Delta\tilde{\mathbf{x}}(t)|$ of a pair of initially nearby trajectories, one can instead solve for the separation in tangent space, denoted by $\delta\tilde{\mathbf{x}}(t)$, and substitute this distance into Eq. (7). The tangent-space separation $\delta\tilde{\mathbf{x}}(t)$ can be dynamically computed by assuming an infinitesimal initial perturbation to a reference trajectory starting at $\tilde{\mathbf{x}}(0) = \tilde{\mathbf{x}}_0$, leading to the system of equations

$$\frac{\mathrm{d}\tilde{\mathbf{x}}}{\mathrm{d}t} = F(\tilde{\mathbf{x}}), \tag{8}$$

$$\frac{\mathrm{d}\mathbf{\Phi}}{\mathrm{d}t} = \mathbf{M}\mathbf{\Phi}. \tag{9}$$

Here, $\mathbf{\Phi}$ is the fundamental matrix and $M_{ij} \equiv \mathrm{d}F_i/\mathrm{d}x_j$. The tangent-space separation with respect to the initial point in phase-space $\tilde{\mathbf{x}}(0) = \tilde{\mathbf{x}}_0$ is extracted by computing $\delta\tilde{\mathbf{x}}(t) \equiv \mathbf{\Phi}\delta\tilde{\mathbf{x}}(0)$ with $\mathbf{\Phi}(0) = \mathbb{I}$.

As we are only interested in the maximum Lyapunov exponent, it suffices to choose the initial separation $\delta\tilde{\mathbf{x}}(0)$ along a random direction in phase-space, and we propagate Eqs. (8) and (9) for each initial condition $\tilde{\mathbf{x}}_0$ for sufficiently large t that our estimate of λ_{L} from Eq. (7) converges.

Connection between classical and quantum Lyapunov exponents. In our discussion of the exponential growth of FOTOCs, we have argued that λ_{Q} is intimately

related to the classical Lyapunov exponent λ_L. Specifically, we have that $\lambda_Q \simeq 2\lambda_L$. Here, we further articulate this connection using a semi-classical description of the quantum dynamics, specifically by considering the evolution in the truncated Wigner approximation (TWA)[39].

First, we remind the reader that for a small perturbation $\delta\phi$, a FOTOC $\mathcal{F}_G(t,\delta\phi)$ can be expanded to $\mathcal{O}(\delta\phi^2)$ as $\mathcal{F}_G(t,\delta\phi) \approx 1 - \delta\phi^2 \mathrm{var}(\hat{G})$. A simple conclusion from this expansion is that if $\mathcal{F}_G(t,\delta\phi)$ grows exponentially we can attribute this behavior to the variance, i.e. it must be true that $\mathrm{var}(\hat{G}) \rangle \sim e^{\lambda_Q t}$.

A semi-classical explanation of this exponential growth is simplified by assuming that $\hat{G}$ is an operator which is linear in the classical phase-space variables $\tilde{\mathbf{x}}$. For concreteness, let us consider $\hat{G} = \hat{X} = \frac{1}{2}(\hat{a} + \hat{a}^\dagger)$ as in Fig. 3 of the main text, which corresponds to α_R in the classical phase-space.

Next, we consider a description of the quantum dynamics within the framework of the TWA. Here, the dynamics is computed by solving the classical equations of motion (Eq. (5)) with random initial conditions sampled from the corresponding Wigner phase-space distribution of the initial state[39]. Quantum expectation values are then obtained by appropriate averaging over an ensemble of trajectories, e.g., $\langle \hat{X} \rangle \equiv \overline{\alpha_R}$ where the overline denotes a stochastic average. The random sampling of initial conditions serves to model the quantum fluctuations of the initial state.

For a classically meaningful initial state (i.e. a product of coherent states for the phonon and spin degrees of freedom), the fluctuations in each of the phase-space variables are typically Gaussian and centered around the expectation values of the initial state. A concrete example to illustrate this is the state $|\Psi_0^c\rangle = |(-N/2)_x\rangle \otimes |0\rangle$ considered in the main text. For each trajectory, the variable $(\alpha_R)_j$ (j denoting the trajectory), for example, is sampled from a Gaussian distribution with mean zero and variance 1/4. The connection between the quantum dynamics and classical chaos is made by instead considering sampling only the fluctuations $\delta\alpha_R$ about a central classical trajectory, i.e. $(\alpha_R)_j \to \alpha_R^{cl} + (\delta\alpha_R)_j$.

Solving the dynamics of the central classical trajectory and the ensemble of fluctuations is then identical to the calculation of Eqs. (8) and (9), from which the Lyapunov exponent is calculated. In particular, the connection between quantum and classical exponents is finally made clear by evaluating the quantum variance,

$$\mathrm{var}(\hat{X}) = \left(\overline{\alpha_R^2} - \overline{\alpha_R}^2\right), \tag{10}$$

$$\equiv \left(\overline{\delta\alpha_R^2} - \overline{\delta\alpha_R}^2\right). \tag{11}$$

As $\delta\alpha_R$ is evaluated directly from Eq. (9), then we expect from our previous calculations that $|\delta\alpha_R| \sim e^{\lambda_L t}$ for a generic random perturbation, sampled according to the TWA prescription, in parameter regimes where there is classical chaos. Thus, we extrapolate that the quantum variance will grow like $\overline{\delta\alpha_R^2} - \overline{\delta\alpha_R}^2 \sim e^{2\lambda_L t}$. Inspection of this final result shows that we should expect $\lambda_Q \simeq 2\lambda_L$.

Connection between FOTOCs and RE. The connection between the FOTOCs and entanglement entropy is best established by first considering the case of the spin−phonon RE $S_2(\hat{\rho}_{\mathrm{ph}})$. We begin by writing the purity of the reduced density matrix explicitly in terms of the elements of the density matrix,

$$\mathrm{Tr}\left[\hat{\rho}_{\mathrm{ph}}^2(t)\right] = \sum_{\substack{n,n' \\ m_r,m_r'}} \rho_{m_r,m_r}^{n,n'}(t)\rho_{m_r',m_r}^{n',n}(t). \tag{12}$$

Our insight is that, in the case of a pure global state, the summation in Eq. (12) for the purity of the reduced density matrix can be manipulated and re-expressed as

$$\mathrm{Tr}\left[\hat{\rho}_{\mathrm{ph}}^2(t)\right] \equiv I_0^{\hat{S}_r}(t) + I_0^{\hat{n}}(t) - D_{\mathrm{diag}}^{\hat{S}_r,\hat{n}}(t) + C_{\mathrm{off}}^{\hat{S}_r,\hat{n}}(t), \tag{13}$$

where

$$D_{\mathrm{diag}}^{\hat{S}_r,\hat{n}}(t) = \sum_{\substack{n, \\ m_r}} \left[\rho_{m_r,m_r}^{n,n}(t)\right]^2, \tag{14}$$

$$C_{\mathrm{off}}^{\hat{S}_r,\hat{n}}(t) = \sum_{\substack{n\neq n' \\ m_r\neq m_r'}} \rho_{m_r,m_r}^{n,n'}(t)\rho_{m_r',m_r}^{n',n}(t), \tag{15}$$

are the sum of the squared diagonal elements of the density matrix and the sum over the off-diagonal coherences, respectively. Thus, we seek to understand when these latter terms can be neglected and thus the purity (and associated entropy) is expressible in terms of only the $I_0^{\hat{G}}$.

Firstly, there is the case of a large transverse field, $B \gg B_c$ and an initial state which is polarized along the direction of the transverse field with vacuum occupation, i.e., $|\Psi_0\rangle = |(\pm N/2)_x\rangle \otimes |0\rangle$. In this case, we expect the collective spin to remain strongly polarized along the field direction. If we choose the FOTOC spin rotation axis to be along that of the initial state and transverse field, $\hat{S}_r = \hat{S}_x$, then we have that $C_{\mathrm{off}}^{\hat{S}_x,\hat{n}}(t) \approx 0$ due to the absence of initial coherences between the spin sectors in this basis, and by similar reasoning $I_0^{\hat{n}}(t) \approx D_{\mathrm{diag}}^{\hat{S}_x,\hat{n}}(t)$. Hence, we expect Eq. (13) to simplify so that $\mathrm{Tr}[\hat{\rho}_{\mathrm{ph}}^2(t)] \approx I_0^{\hat{S}_x}(t)$. Identical reasoning can be applied in the normal phase ($B > B_c$) when the phonon detuning is the largest energy scale, such that $\mathrm{Tr}[\hat{\rho}_{\mathrm{ph}}^2(t)] \approx I_0^{\hat{n}}(t)$.

The second scenario is closely related to the first. Consider an initial coherent spin state polarized along an arbitrary spin direction and vacuum phonon occupation. For arbitrary transverse field strength and on sufficiently short time-scales $t \lesssim \lambda_Q^{-1}$, then the spin component of the evolved state remains largely polarized along a particular axis dictated by the initial state. Similar to the first scenario, by choosing the spin rotation of the FOTOC, $\hat{S}_r$, to match the polarization of the initial state, then we will have $\mathrm{Tr}[\hat{\rho}_{\mathrm{ph}}^2(t)] \approx I_0^{\hat{S}_r}(t)$. This is justified as $C_{\mathrm{off}}^{\hat{S}_r,\hat{n}}(t)$ and $I_0^{\hat{n}}(t) + D_{\mathrm{diag}}^{\hat{S}_r,\hat{n}}(t)$ vanish, as again the state at short times will not have appreciable coherences between different spin sectors in this basis.

Lastly, for a small transverse field, $B \ll B_c$, and beyond short times $t \gtrsim \lambda_Q^{-1}$ (i.e., beyond the time-scale when the spin state is still strongly polarized and the second scenario is still valid), we expect Eq. (13) to be well approximated by $\mathrm{Tr}[\hat{\rho}_{\mathrm{ph}}^2(t)] \approx I_0^{\hat{S}_r}(t) + I_0^{\hat{n}}(t)$ for any spin rotation axis $\hat{S}_r$. This is because initially pure states which are sufficiently scrambled after a quench of the system parameters closely resemble so-called canonical pure thermal quantum (cTPQ) states[60] in a generic basis. For cTPQ states, the summation over off-diagonal coherences $C_{\mathrm{off}}^{\hat{S}_r,\hat{n}}$ vanishes exactly for a sufficiently large system as the coherences can be considered as random variables[60]. Moreover, for a typical spin rotation axis $\hat{S}_r$, the cTPQ state will have a spin distribution $P(M_{\hat{S}_r})$ which is largely delocalized implying that $D_{\mathrm{diag}}^{\hat{S}_r,\hat{n}} \sim 1/(N n_{\mathrm{ph}})$ where n_{ph} is some constant which characterizes the spread of the boson number distribution. The term $D_{\mathrm{diag}}^{\hat{S}_r,\hat{n}}$ is then typically much smaller in magnitude when compared to the remaining terms $I_0^{\hat{S}_r}$ and $I_0^{\hat{n}}$. This reasoning leads to $\mathrm{Tr}[\hat{\rho}_{\mathrm{ph}}^2(t)] \approx I_0^{\hat{S}_r}(t) + I_0^{\hat{n}}(t)$. Discussion of the sensitivity of these arguments to the rotation direction can be found in Supplementary Methods.

More generally, we can extend these arguments to extract a correspondence with the Renyi entropy of a generic bipartition of the spin−phonon system. Specifically, splitting the system $\mathcal{S}$ into a subsystem A: L spin-1/2s, and its complement A_c: $N - L$ spin-1/2s and the bosonic mode. In the weak-field regime $B \ll B_c$, $\mathrm{Tr}[\hat{\rho}_A^2] \approx I_0^A + I_0^{A_c}$. Here, the terms I_0^A and $I_0^{A_c}$ are obtained as the Fourier amplitudes of fidelity OTOCs for generalized rotations within each subsystem. Specifically, a local rotation $e^{i\phi\hat{S}_{r,A}}$ taken to act on the spin-12s in the A subsystem, and a joint (but uncorrelated) rotation $e^{i\phi\hat{S}_{r,A_c}} e^{i\theta\hat{a}^\dagger\hat{a}}$ of the spins and bosons in the complement A_c.

Experimental implementation. By preparing an initial spin polarized state, recent experiments[18] demonstrated it was possible to measure the many-body overlap of the final state with the initial configuration by fluorescence detection. The Dicke model, however, includes spin and phonon degrees of freedom.

While the full spin–phonon fidelity measurement has not yet been demonstrated experimentally, such measurement is possible by extending the method in[55] to a multi-qubit system. In particular, we note that this proposal is comprised of a two-step measurement, where we first measure the spin degree of freedom. The probability of all ions being in the dark state (i.e. all in the state $|\downarrow\rangle_z$) can be measured with excellent fidelity and has been previously demonstrated[18]. The dark state does not scatter photons, and as such, this measurement will not change the state of the phonons. Next one can proceed to measure the phonon occupation via the protocol described in ref. [55].

Finally, as noted in the main text, we have taken into account the single-particle decoherence present in the experiment. The results presented in the main text accounted for this by approximating the effects of decoherence by an exponential decay, $\bar{I}_0^{\hat{G}} \to \bar{I}_0^{\hat{G}} e^{-\Gamma N t}$. We have justified this approximation by comparing to an efficient numerical solution of the full Lindblad master equation[19,61,62] for smaller system sizes ($N = 10$). We find that the decoherence is well-captured by the approximate model for all transverse field strengths B considered.

Thermal and diagonal ensembles. The canonical thermal ensemble, used in Fig. 4, is defined by the density matrix $\hat{\rho}_{\mathrm{therm}} = e^{-\beta\hat{H}_D}/\mathrm{Tr}[e^{-\beta\hat{H}_D}]$, which is characterized by the inverse temperature $\beta = 1/(k_B T)$. This inverse temperature is chosen such that energy of the ensemble is matched to that of the initial state of the dynamics, $\langle E \rangle_{\mathrm{therm}} \equiv \mathrm{Tr}[\hat{H}_D\hat{\rho}_{\mathrm{therm}}] = \langle\Psi_0|\hat{H}_D|\Psi_0\rangle$. The RE for bipartitions of the thermal ensemble is then obtained via the definition $S_2^{\mathrm{therm}} \equiv -\log\left(\mathrm{Tr}[(\hat{\rho}_{L_A}^{\mathrm{therm}})^2]\right)$ where $\hat{\rho}_{L_A}^{\mathrm{therm}} = \mathrm{Tr}_{\mathrm{ph},N-L_A}(\hat{\rho}_{\mathrm{therm}})$ is the reduced density matrix obtained after tracing out the phonon degree of freedom and the remaining $N - L_A$ spins.

A related concept is the diagonal ensemble $\hat{\rho}_D$[1,63], which generically describes the (time-averaged) observables of a quantum system which has relaxed at long times. The ensemble is defined as the mixed state $\hat{\rho}_D \equiv \sum_{E_n} |c_{E_n}|^2 |E_n\rangle\langle E_n|$, where $c_{E_n} \equiv \langle\Psi_0|E_n\rangle$ and $|E_n\rangle$ are the eigenstates of the Hamiltonian $\hat{H}_D$ with associated

eigenvalue E_n. We use this diagonal ensemble as a comparison to the time-averaged distribution functions $P(M_z)$ and $P(n)$ in Fig. 4.

Data availability

The source data underlying Figs. 2–4 of the main text are provided as a source data file. Additional numerical data and computer codes used in this study are available from the corresponding author upon request.

Received: 24 September 2018 Accepted: 5 March 2019
Published online: 05 April 2019

References

1. D'Alessio, L., Kafri, Y., Polkovnikov, A. & Rigol, M. From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics. *Adv. Phys.* **65**, 239–362 (2016).
2. Nandkishore, R. & Huse, D. A. Many-body localization and thermalization in quantum statistical mechanics. *Annu. Rev. Condens. Matter Phys.* **6**, 15–38 (2015).
3. Gogolin, C. & Eisert, J. Equilibration, thermalisation, and the emergence of statistical mechanics in closed quantum systems. *Rep. Progress Phys.* **79**, 056001 (2016).
4. Kaufman, A. M. et al. Quantum thermalization through entanglement in an isolated many-body system. *Science* **353**, 794–800 (2016).
5. Neill, C. et al. Ergodic dynamics and thermalization in an isolated quantum system. *Nat. Phys.* **12**, 1037–1041 (2016).
6. Clos, G., Porras, D., Warring, U. & Schaetz, T. Time-resolved observation of thermalization in an isolated quantum system. *Phys. Rev. Lett.* **117**, 170401 (2016).
7. Brydges, T. et al. Probing entanglement entropy via randomized measurements Preprint at https://arxiv.org/abs/1806.05747 (2018).
8. Hayden, P. & Preskill, J. Black holes as mirrors: quantum information in random subsystems. *J. High Energy Phys.* **2007**, 120 (2007).
9. Sekino, Y. & Susskind, L. Fast scramblers. *J. High Energy Phys.* **2008**, 065 (2008).
10. Shenker, S. H. & Stanford, D. Black holes and the butterfly effect. *J. High Energy Phys.* **2014**, 067 (2014).
11. Hosur, P., Qi, X. L., Roberts, D. A. & Yoshida, B. Chaos in quantum channels. *J. High Energy Phys.* **2016**, 004 (2016).
12. Kitaev, A. in *Talk at Fundamental Physics Prize Symposium* at Stanford University, November 10 (2014).
13. Swingle, B., Bentsen, G., Schleier-Smith, M. & Hayden, P. Measuring the scrambling of quantum information. *Phys. Rev. A* **94**, 040302 (2016).
14. Maldacena, J. & Stanford, D. Remarks on the sachdev-ye-kitaev model. *Phys. Rev. D* **94**, 106002 (2016).
15. Yao, N. Y. et al. Interferometric approach to probing fast scrambling. Preprint at https://arxiv.org/abs/1607.01801 (2016).
16. Shen, H., Zhang, P., Fan, R. & Zhai, H. Out-of-time-order correlation at a quantum phase transition. *Phys. Rev. B* **96**, 054503 (2017).
17. Zhu, G., Hafezi, M. & Grover, T. Measurement of many-body chaos using a quantum clock. *Phys. Rev. A* **94**, 062329 (2016).
18. Gärttner, M. et al. Measuring out-of-time-order correlations and multiple quantum spectra in a trapped ion quantum magnet. *Nat. Phys.* **13**, 781–786 (2017).
19. Gärttner, M., Hauke, P. & Rey, A. M. Relating out-of-time-order correlations to entanglement via multiple-quantum coherences. *Phys. Rev. Lett.* **120**, 040402 (2018).
20. Li, J. et al. Measuring out-of-time-order correlators on a nuclear magnetic resonance quantum simulator. *Phys. Rev. X* **7**, 031011 (2017).
21. Meier, E. J., Ang'ong'a, J., An, F. A. & Gadway, B. Exploring quantum signatures of chaos on a floquet synthetic lattice. Preprint at https://arxiv.org/abs/1705.06714 (2017).
22. Dicke, R. H. Coherence in spontaneous radiation processes. *Phys. Rev.* **93**, 99–110 (1954).
23. Baumann, K., Guerlin, C., Brennecke, F. & Esslinger, T. Dicke quantum phase transition with a superfluid gas in an optical cavity. *Nature* **464**, 1301–1306 (2010).
24. Baumann, K., Mottl, R., Brennecke, F. & Esslinger, T. Exploring symmetry breaking at the dicke quantum phase transition. *Phys. Rev. Lett.* **107**, 140402 (2011).
25. Klinder, J., Keßler, H., Wolke, M., Mathey, L. & Hemmerich, A. Dynamical phase transition in the open dicke model. *Proc. Natl. Acad. Sci. USA* **112**, 3290–3295 (2015).
26. Zhang, Z. et al. Dicke-model simulation via cavity-assisted raman transitions. *Phys. Rev. A* **97**, 043858 (2018).
27. Safavi-Naini, A. et al. Verification of a many-ion simulator of the dicke model through slow quenches across a phase transition. *Phys. Rev. Lett.* **121**, 040503 (2018).
28. Pérez-Fernández, P. et al. Excited-state phase transition and onset of chaos in quantum optical models. *Phys. Rev. E* **83**, 046208 (2011).
29. Brandes, T. Excited-state quantum phase transitions in dicke superradiance models. *Phys. Rev. E* **88**, 032133 (2013).
30. Emary, C. & Brandes, T. Chaos and the quantum phase transition in the dicke model. *Phys. Rev. E* **67**, 066203 (2003).
31. Buijsman, W., Gritsev, V. & Sprik, R. Nonergodicity in the anisotropic dicke model. *Phys. Rev. Lett.* **118**, 080601 (2017).
32. Altland, A. & Haake, F. Quantum chaos and effective thermalization. *Phys. Rev. Lett.* **108**, 073601 (2012).
33. Altland, A. & Haake, F. Equilibration and macroscopic quantum fluctuations in the dicke model. *New J. Phys.* **14**, 073011 (2012).
34. Emary, C. & Brandes, T. Quantum chaos triggered by precursors of a quantum phase transition: the Dicke model. *Phys. Rev. Lett.* **90**, 044101 (2003).
35. Chávez-Carlos, J., Bastarrachea-Magnani, M. A., Lerma-Hernández, S. & Hirsch, J. G. Classical chaos in atom-field systems. *Phys. Rev. E* **94**, 022209 (2016).
36. Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, 2nd edn (Westview Press, Boulder, CO, USA, 2015).
37. Schmitt, M., Sels, D., Kehrein, S. & Polkovnikov, A. Semiclassical echo dynamics in the sachdev-ye-kitaev model. Preprint at https://arxiv.org/abs/1802.06796 (2018).
38. Fox, R. F. & Elston, T. C. Chaos and a quantum-classical correspondence in the kicked top. *Phys. Rev. E* **50**, 2553–2563 (1994).
39. Polkovnikov, A. Phase space representation of quantum dynamics. *Ann. Phys.* **325**, 1790–1852 (2010).
40. Schachenmayer, J., Pikovski, A. & Rey, A. M. Many-body quantum spin dynamics with monte carlo trajectories on a discrete phase space. *Phys. Rev. X* **5**, 011022 (2015).
41. Toth, G. & Apellaniz, I. Quantum metrology from a quantum information science perspective. *J. Phys. A Math. Theor.* **47**, 424006 (2014).
42. Gessner, M., Pezzé, L. & Smerzi, A. Efficient entanglement criteria for discrete, continuous, and hybrid variables. *Phys. Rev. A* **94**, 020101 (2016).
43. Macrì, T., Smerzi, A. & Pezzè, L. Loschmidt echo for quantum metrology. *Phys. Rev. A* **94**, 010102 (2016).
44. Pezzé, L. & Smerzi, A. Entanglement, nonlinear dynamics, and the Heisenberg Limit. *Phys. Rev. Lett.* **102**, 100401 (2009).
45. Hyllus, P. et al. Fisher information and multiparticle entanglement. *Phys. Rev. A* **85**, 022321 (2012).
46. Tóth, G. Multipartite entanglement and high-precision metrology. *Phys. Rev. A* **85**, 022322 (2012).
47. Rozenbaum, E. B., Ganeshan, S. & Galitski, V. Lyapunov exponent and out-of-time-ordered correlator's growth rate in a chaotic system. *Phys. Rev. Lett.* **118**, 086801 (2017).
48. Baum, J., Munowitz, M., Garroway, A. N. & Pines, A. Multiple-quantum dynamics in solid state nmr. *J. Chem. Phys.* **83**, 2015–2025 (1985).
49. Alvarez, G. A., Suter, D. & Kaiser, R. Localization−delocalization transition in the dynamics of dipolar-coupled nuclear spins. *Science* **349**, 846–848 (2015).
50. Sánchez, C. M., Levstein, P. R., Acosta, R. H. & Chattah, A. K. Nmr loschmidt echoes as quantifiers of decoherence in interacting spin systems. *Phys. Rev. A* **80**, 012328 (2009).
51. Wall, M. L., Safavi-Naini, A. & Rey, A. M. Boson-mediated quantum spin simulators in transverse fields: *xy* model and spin-boson entanglement. *Phys. Rev. A* **95**, 013602 (2017).
52. Jurcevic, P. et al. Direct observation of dynamical quantum phase transitions in an interacting many-body system. *Phys. Rev. Lett.* **119**, 080501 (2017).
53. Zhang, J. et al. Observation of a many-body dynamical phase transition with a 53-qubit quantum simulator. *Nature* **551**, 601–604 (2017).
54. Cohn, J. et al. Bang-bang shortcut to adiabaticity in the dicke model as realized in a penning trap experiment. *New J. Phys.* **20**, 055013 (2018).
55. Gebert, F., Wan, Y., Wolf, F., Christoph Help, J. & Schmidt, P. O. Corrigendum: detection of motional ground state population using delayed pulses. *New J. Phys.* **20**, 029501 (2018).
56. Ge, W. et al. Trapped ion quantum information processing with squeezed phonons. *Phys. Rev. Lett.* **122**, 030501 (2019).
57. Alavirad, Y. & Lavasani, A. Scrambling in the Dicke model. Preprint at https://arxiv.org/abs/1808.02038 (2018).
58. Chávez-Carlos, J. et al. Quantum and classical lyapunov exponents in atom-field interaction systems. *Phys. Rev. Lett.* **122**, 024101 (2019).
59. Skokos, Ch. The Lyapunov characteristic exponents and their computation. In *Dynamics of* Small Solar System Bodies *and Exoplanets* (eds Souchay, J. J. & Dvorak, R.) 63–135 (Springer, New York, 2010).
60. Nakagawa, Y. O., Watanabe, M., Fujita, H. & Sugiura, S. Universality in volume-law entanglement of scrambled pure quantum states. *Nat. Commun.* **9**, 1635 (2018).

61. Xu, M., Tieri, D. A. & Holland, M. J. Simulating open quantum systems by applying su(4) to quantum master equations. *Phys. Rev. A* **87**, 062101 (2013).
62. Shammah, N., Ahmed, S., Lambert, N., De Liberato, S. & Nori, F. Open quantum systems with local and collective incoherent processes: efficient numerical simulations using permutational invariance. *Phys. Rev. A* **98**, 063815 (2018).
63. Rigol, M., Dunjko, V. & Olshanii, M. Thermalization and its mechanism for generic isolated quantum systems. *Nature* **452**, 854–858 (2008).

Acknowledgements

We thank A. Kaufman and R. Nandkishore for fruitful discussions. This work is supported by the Air Force Office of Scientific Research grants FA9550-18-1-0319 and its Multidisciplinary University Research Initiative grant (MURI), by the Defense Advanced Research Projects Agency (DARPA) and Army Research Office grant W911NF-16-1-0576, and W911NF-19-1-0210, the National Science Foundation grant PHY-1820885, JILA-NSF grant PFC-173400, and the National Institute of Standards and Technology.

Author contributions

The calculations were performed by R.J.L.-S. and A.S.-N. All authors (R.J.L.-S., A.S.-N., A.M.R. and J.J.B.) participated in the conception of the project, analysis of the results and preparation of the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09436-y.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

npj | Quantum Information

www.nature.com/npjqi

ARTICLE OPEN

Preserving entanglement during weak measurement demonstrated with a violation of the Bell–Leggett–Garg inequality

TC White[1,4,5], JY Mutus[1,4,5], J Dressel[2], J Kelly[1], R Barends[1,5], E Jeffrey[1,5], D Sank[1,5], A Megrant[1,3], B Campbell[1], Yu Chen[1,5], Z Chen[1], B Chiaro[1], A Dunsworth[1], I-C Hoi[1], C Neill[1], PJJ O'Malley[1], P Roushan[1,5], A Vainsencher[1], J Wenner[1], AN Korotkov[2] and John M Martinis[1,5]

Weak measurement has provided new insight into the nature of quantum measurement, by demonstrating the ability to extract average state information without fully projecting the system. For single-qubit measurements, this partial projection has been demonstrated with violations of the Leggett–Garg inequality. Here we investigate the effects of weak measurement on a maximally entangled Bell state through application of the Hybrid Bell–Leggett–Garg inequality (BLGI) on a linear chain of four transmon qubits. By correlating the results of weak ancilla measurements with subsequent projective readout, we achieve a violation of the BLGI with 27 s.d.s. of certainty.

npj Quantum Information (2016) **2,** 15022; doi:10.1038/npjqi.2015.22; published online 16 February 2016

INTRODUCTION

Quantum computing promises greater processing power through the clever application of superposition and entanglement. Despite the importance of this uniquely quantum behaviour, it occurs elusively behind the non-unitary effects of measurement collapse. Weak measurements[1–3] have provided new insight into this process by demonstrating the ability to extract average state information without fully collapsing the system. These gentler measurements have allowed single-configuration violations of the Leggett–Garg inequality[4–11] and, more trecently, the detailed tracking of single-qubit trajectories.[12,13] It is an outstanding challenge, however, to achieve the same level of measurement control with an entangled state. Here we demonstrate a continuous and controlled exchange between extracted single-qubit state information and two-qubit entanglement collapse, through the unique framework of the Bell–Leggett–Garg inequality (BLGI). We quantify this effect by correlating variable strength ancilla qubit measurements with subsequent projective readout to collect all the statistics of a Bell inequality experiment[14–17] in a single quantum circuit. In addition, we demonstrate the ability to measure the Bell state with minimal entanglement collapse, by violating this hybrid BLGI[18] at the weakest measurement strengths. This experiment indicates that it is possible carry out high-fidelity ancilla measurement in large entangled systems. In addition, combining this experiment with remote entanglement methods[19] may eventually lead to a loophole-free violation of classical hidden variable theories.

The challenge of successfully implementing weak measurements is twofold: the first challenge is to evaluate the amount of information extracted on average by the measurement; the second challenge is to evaluate the measurement back-action on the system. For a single-qubit state, the Leggett–Garg inequality[20] (LGI) provides an elegant way to do both with a single experiment. The LGI was originally designed to verify the 'quantumness' of macroscopic objects through the effects of projective measurement, which allows larger correlations between successive measurements (e.g., at times $t_1 < t_2 < t_3$) than are possible classically. More recent generalisations of the LGI prepare a known state at time t_1 and replace the intermediate measurement at time t_2 with a weak measurement.[4–11] This minimises the quantum-state disturbance while still extracting sufficient information on average. Ideally, this allows all the statistics necessary for a violation of the inequality to be measured with a single experimental configuration. Violating the inequality in this way guarantees that the state information has been extracted without significant back-action on the system.[7]

Evaluating the effect of weak measurements on an entangled state is more difficult because the degree of entanglement is generally challenging to quantify. The most robust method for quantifying entanglement remains a Bell test,[14] which was first proposed by Bell and later refined by Clauser, Horne, Shimony and Holt into an inequality (CHSH). The CHSH term sums the correlation measurements of two spatially separated qubits in four different measurement bases and bounds the maximum total value of classical correlations to be $|\mathrm{CHSH}|_{\mathrm{class}} \leqslant 2$.

In superconducting qubits, we use qubit state rotations to map the desired measurement basis onto the ground ($|0\rangle$) and excited ($|1\rangle$) states of the system. For measurement rotations a (qubit 1) and b (qubit 2), the correlation amplitude between two

[1]Department of Physics, University of California, Santa Barbara, CA, USA; [2]Department of Electrical and Computer Engineering, University of California, Riverside, CA, USA and [3]Department of Materials, University of California, Santa Barbara, CA, USA.
Correspondence: JM Martinis (martinis@physics.ucsb.edu)
[4]These authors contributed equally to this work.
[5]Current address: Google Inc., Santa Barbara, CA 93117, USA.
Received 30 July 2015; revised 6 October 2015; accepted 30 October 2015

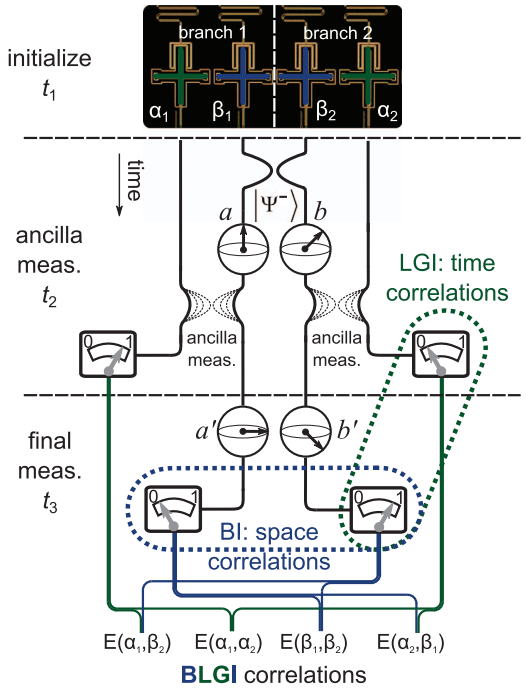


Figure 1. Schematic of the hybrid Bell–Leggett–Garg inequality and optical micrograph of the superconducting quantum device. The algorithm consists of two LGI weak measurement branches, bridged by the entanglement of the central Bell qubits. The Bell pair ($\beta_{1,2}$) is initially prepared in the anti-symmetric singlet Bell state $|\Psi^-\rangle$. Next, each Bell qubit is rotated to its first measurement basis (a or b) and entangled with its ancilla qubit ($\alpha_{1,2}$). Finally, the Bell qubits are rotated and projectively read out in bases corresponding to angles a' and b'. By correlating the final projective readout and the weak ancilla measurements, we calculate all four terms of a CHSH correlator simultaneously.

measurements is given by

$$E(a,b) = P(00) - P(10) - P(01) + P(11), \quad (1)$$

where the term $P(00)$ is the probability that both qubits are in the ground state. A traditional CHSH experiment combines the four correlator terms

$$\mathrm{CHSH} = E(a,b) + E(a',b) + E(a,b') - E(a',b'). \quad (2)$$

Entangled quantum states can violate the classical bound, with a fully entangled Bell state ideally saturating the quantum upper bound of $|\mathrm{CHSH}|_{\mathrm{quant}} \leqslant 2\sqrt{2}$ at the specific rotation angles $a = 0$, $b = \pi/4$, $a' = \pi/2$ and $b' = 3\pi/4$.

To understand the effect of weak measurement on an entangled state, we combine the spatial correlations of a Bell inequality with the temporal correlations of an LGI to construct a BLGI. The algorithm, as described by Dressel and Korotkov[18] and shown in Figure 1, consists of a CHSH-style experiment in which each Bell qubit is measured twice in succession as for a simultaneous LGI.[7,10,27] The initial measurements are carried out by ancilla qubits, which act as probes of the entangled system before being projectively read out. By varying the degree of entanglement between each ancilla and the Bell qubit it probes, we can vary the strength of the measurement. After preparing the Bell qubits in the anti-symmetric singlet state

$$|\Psi^-\rangle = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle), \quad (3)$$

each Bell qubit ($\beta_{1,2}$) is rotated to its first measurement angle ($a = 0$, $b = \pi/4$) and then entangled with its ancilla qubit ($\alpha_{1,2}$) to implement the tunable-strength measurement. Next, each Bell qubit is rotated to its final measurement angle ($a' = \pi/2$, $b' = 3\pi/4$), and all four qubits are read out. With this procedure, the data for each measurement angle are encoded on a distinct qubit ($a \to \alpha_1$, $b \to \alpha_2$, $a' \to \beta_1$, and $b' \to \beta_2$). The BLGI correlator then takes the form similar to equation (2)

$$\langle C \rangle = -E(\alpha_1,\alpha_2) - E(\alpha_1,\beta_2) + E(\beta_1,\alpha_2) - E(\beta_1,\beta_2). \quad (4)$$

where each term is calculated as in equation (1).

The BLGI bounds correlations between these four distinct measurements based on the classical assumptions of local-macro-realism[18] (See Supplementary Information at link for further discussion of the sample, measurement techniques and mathematical assumptions). Classically, the measurement of one qubit in the Bell pair should have no affect on the other, and the strength of the ancilla measurement should have no effect on the result of the following projective readout. By encoding the measurement result for each rotation angle on an independent qubit, we can test both of these assumptions at the same time. Quantum mechanically, if the ancilla measurement is fully projective then we should measure the expected correlation amplitude of $|E(\alpha_1,\alpha_2)| = -1/\sqrt{2}$ only in the initial measurement basis, as the Bell qubits are no longer entangled after that measurement. As the ancilla measurement strength is decreased, we should extract the same qubit information on average while only partially collapsing the Bell state. For sufficiently weak measurements, the magnitudes of all four correlators should approach the unperturbed Bell state values of $1/\sqrt{2}$ while $\langle C \rangle$ approaches $2\sqrt{2}$. Thus, a violation implies that our system demonstrates non-classical correlations through the entanglement of the Bell pair, while also demonstrating the ability to extract average state information without significant back-action on the entanglement of the system.

RESULTS

To quantify the entanglement collapse, we measure each two-qubit correlator in $\langle C \rangle$ versus ancilla measurement strength ϕ. The data are plotted in Figure 2 alongside theory curves generated by a quantum model that includes realistic environmental dephasing and readout fidelity.[18] Error bars for the data represent ± 10 s.d. of the mean to demonstrate the increase in noise with decreasing measurement strength. For projective angles $\phi \approx \pi/2$, the ancilla measurement results ($E(\alpha_1, \alpha_2)$) reflect the correlation expected from a fully collapsed Bell pair. As the measurement strength is decreased, this ancilla correlation remains nearly constant while additional inter-qubit correlations ($E(\alpha_1, \beta_2)$, $E(\beta_1, \alpha_2)$, $E(\beta_1, \beta_2)$) emerge. For sufficiently weak measurements, $\langle C \rangle$ exceeds the classical bound of 2 and saturates towards the CHSH value of 2.5, which is expected from simulations for a fully entangled Bell state in realistic experimental conditions (see Supplementary Information at link for further discussion of the sample, measurement techniques and mathematical assumptions).

The measured BLGI correlations follow the theoretical model very closely for all measurement strengths (see Supplementary Information at link for further discussion of the sample, measurement techniques, and mathematical assumptions). This behaviour reveals the continuous and controlled exchange between the collapse of an entangled Bell state and the single-qubit state information gained from tunable-strength measurements. Each ancilla qubit, when calibrated, retains the same correlations for all measurement strengths, whereas each Bell qubit has its correlations damped through partial projection by its ancilla qubit.[18] The effect of partial projection can be seen in the difference in functional behaviour between the Bell-ancilla ($E(\alpha_1, \beta_2)$, $E(\beta_1, \alpha_2)$) and the Bell–Bell ($E(\beta_1, \beta_2)$) correlator terms. In the Bell-ancilla terms, the correlations are suppressed solely because of the randomisation of the Bell qubit,

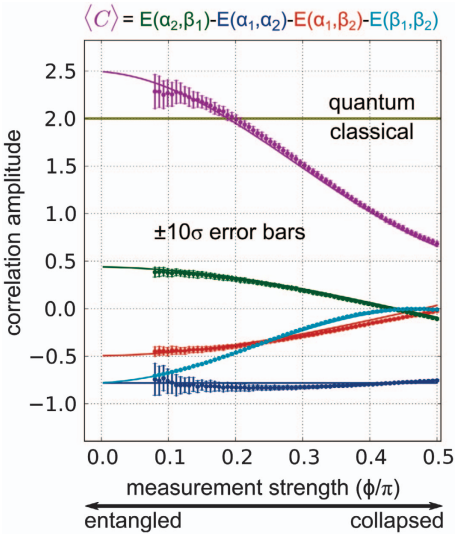


Figure 2. Graph showing both experimental data (points) and theoretical predictions (lines) for the correlator $\langle C \rangle$ and its four terms versus measurement strength ϕ. The horizontal gold line denotes the classical bound on $\langle C \rangle$. The data set was taken by averaging together 200 traces in which each point was measured 3,000 times for a total of 600,000 iterations per point. The error bars represent 10 s.d. of the mean to demonstrate the scaling of the ancilla measurement noise versus ϕ. The magnitude of the correlations between each pair of qubits reveals the extent to which entanglement has been broken for each measurement strength.

but they return as soon as measurement strength is decreased. In the Bell–Bell term, this effect is compounded, as both qubits are being damped by partial ancilla projection, and thus the correlations return more slowly. This gives $E(\beta_1, \beta_2)$ its distinct shape compared with the other correlators. The BLGI terms can be seen in greater detail in the Supplementary Information (see Supplementary Information at link for further discussion of the sample, measurement techniques and mathematical assumptions).

DISCUSSION

Although other methods exist to characterise entanglement, the correlation measurements of a CHSH experiment remain one of the more robust tests of quantum behaviour because of the considerations given to experimental loopholes. Fortunately, the unique construction of the BLGI allows us to avoid some of the more pervasive loopholes appearing in traditional Bell or LG inequalities. The simultaneous measurement of all four CHSH terms in a single circuit allows us to avoid any configuration-dependent bias, such as the disjoint sampling loophole.[21] The near unit detection efficiency in superconducting systems (See Supplementary Information at link for further discussion of the sample, measurement techniques, and mathematical assumptions)[22] similarly bypasses the fair sampling loophole,[23] which has hindered the investigation of related hybrid inequalities in optical systems.[9,24] In addition, as the data from each ancilla qubit are only correlated with the data from the Bell and ancilla qubits on the remote LGI branch, we can substantially relax the usual LGI noninvasive measurement assumption to the standard locality assumption needed for a Bell inequality instead.

This locality assumption, fundamental to any Bell inequality, presumes that no classical interactions between remote qubits occur during the correlation measurements. The close proximity of adjacent superconducting qubits on a chip implies that such an interaction cannot be ruled out here. Thus, behaviour that appears to be quantum could, at least in principle, be the result of a fast classical interaction between hidden variables in the system. Although we cannot yet completely rule out these local interactions, there are several promising approaches to closing this loophole as well. The assumptions of the BLGI requires only spatial separation of the central Bell qubits. The speed and fidelity of operations in a superconducting qubit system makes modest spatial separation sufficient, and we can sacrifice some Bell state preparation fidelity to achieve it. Techniques such as remote entanglement through measurement,[19] may soon provide the spatial separation necessary to conduct a loophole-free BLGI experiment.

Despite the few remaining loopholes, the excellent agreement between data and theoretical predictions in this experiment allows us to draw certain likely conclusions about the application of ancilla measurement in superconducting circuit systems. The functional dependence of the BLGI correlator on measurement strength implies that the back-action is dominated by the projectiveness of the measurement. This makes it unlikely that there is some poorly understood classical error mechanism that would make it difficult to implement error correction schemes such as the surface code,[25] which rely on sequential high-fidelity ancilla measurements to detect errors.[26] The violation of the BLGI at the weakest measurement strength also implies that it is possible to extract average state information without significant back-action. Normally, the usefulness of this type of measurement is limited by the large number of statistics required to average the noisy detector output and the coherence time of the qubits limiting the total number of measurements. It should be possible, however, to integrate weak ancilla measurements into a large surface code cell, which would correct for most errors while also allowing for a larger ensemble to be collected from the weak measurement. Thus, weak ancilla measurement could become an important tool in understanding the dynamics of large quantum systems.

In the course of this work, we have demonstrated the continuous and controlled collapse of an entangled state based on the strength of tunable ancilla measurements. This behaviour was quantified using the simultaneous correlation measurements that make up the Bell–Leggett–Garg inequality. The violation of this inequality at the weakest measurement strengths demonstrates the viability of using weak ancilla measurements to conduct many sequential measurements of entangled states. This provides a window into the evolution of entangled states, which is a critical component in scaling to larger quantum systems. With the inclusion of new remote entanglement algorithms, the BLGI may also lead to loophole-free violations of classical hidden variable theories. Last, this demonstrates that as we scale to larger multi-qubit systems, with the fidelity and control achieved here, we gain greater access to the rich physics at the heart of quantum mechanics.

MATERIALS AND METHODS

We performed this experiment on a linear chain of Xmon transmon qubits, shown at the top of Figure 1, with ground to excited state transition frequencies in the 4–6 GHz range.[26] Each qubit is individually addressed with a microwave control line, which can be used for single-qubit *X* or *Y* gates, as well as a DC line for implementing *Z*-gates and frequency control. These control lines are used in conjunction to execute high-fidelity two-qubit gates[28] for entanglement and ancilla measurement. The state of each qubit is measured independently using the dispersive shift of a dedicated readout resonator. Resonators are frequency multiplexed[29] and read out with a broadband parametric amplifier,[30] which allows for fast high-fidelity measurement. Further details of this device can be found in ref. 26 (See Supplementary Information at link for further discussion of the sample, measurement techniques and mathematical assumptions).

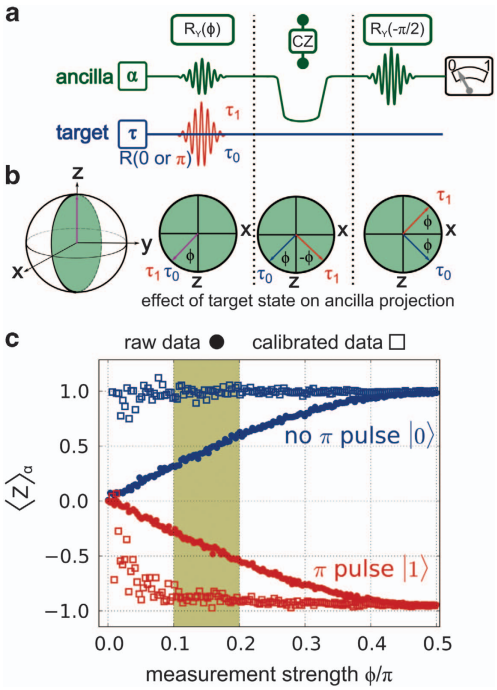


Figure 3. Weak measurement protocol. (**a**) Full pulse sequence of the ancilla measurement algorithm used in the BLGI experiment. The measurement consists of a variable-amplitude Y rotation by an angle ϕ, which controls the strength of the measurement. This is followed by a CZ gate that entangles the ancilla qubit with the target qubit. Finally, the ancilla is rotated by an angle $-\pi/2$, bringing it into the desired measurement basis. Two cases are compared: that of the target qubit in the ground (blue) or excited (red, π rotation) state. (**b**) Bloch sphere representation of the ancilla qubit during the weak measurement protocol when the target qubit is in either the ground (blue) or excited (red) state. The Z averages of the ancilla and target qubit are correlated such that $\langle Z \rangle_a = \sin(\phi)\langle Z \rangle_\tau$, where a full projective measurement corresponds to $\phi = \pi/2$ and no measurement corresponds to $\phi = 0$. (**c**) Ancilla measurement of prepared target state before and after calibrating for measurement strength. We calibrate both curves by the scaling factor required to normalise the average 0 state curve. This is almost equivalent to dividing by sin (ϕ) but bounds the calibrated mean by ± 1. In the calibrated case, the measured mean remains unchanged while the measured variance increases as ϕ decreases. The gold-shaded region denotes angles at which weak measurement data can violate the BLGI while still being reliably calibrated.

The ancilla measurement protocol used in this experiment and shown in Figure 3 is a modified version of the protocol demonstrated in an LGI violation from Groen *et al.*[10] Initially, an ancilla qubit is Y-rotated by an angle $0 \leqslant \phi \leqslant \pi/2$ from its ground state to set the measurement strength. A control phase gate is then performed, causing a Z rotation of $\pi/2$ in the ancilla qubit depending on the target qubit's state. Finally, a $-Y$ rotation of $\pi/2$ is performed on the ancilla qubit to rotate into the correct measurement basis. The visibility of this measurement is then proportional to the distance of the ancilla state vector from the equator of the Bloch sphere, as shown in Figure 3b. When $\phi = \pi/2$, this operation becomes a control-NOT gate and implements a projective measurement. As $\phi \to 0$ the ancilla states become degenerate and no information is extracted.

As the final position of the ancilla state is dependent on the measurement strength, the ancilla readout is imperfectly correlated with the target qubit. That is, the visibility of an ancilla Z average, $\langle Z \rangle_a \simeq \sin(\phi)\langle Z \rangle_\tau$, is compressed from the target Z average by a factor of approximately sin (ϕ). To reconstruct the target Z average from the ancilla Z average, we should thus rescale the signal by 1/sin (ϕ). Initially, this was done by a linear fit of the rotation angle ϕ to the qubit drive amplitude. Unfortunately, this linear fit was too rough to properly calibrate the smallest drive amplitudes and led to a systemic offset in the calibration of $\langle Z \rangle$. To keep the model as simple as possible, we kept the linear fit but used a data-based rescaling to set the measured ground state ($|0\rangle$) average to 1, which ensures that the calibrated ancilla average is properly bounded by ± 1, as shown in Figure 3c. Further details of this calibration can be found in the Supplementary Information (See Supplementary Information at link for further discussion of the sample, measurement techniques and mathematical assumptions).

ACKNOWLEDGEMENTS

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Office grant W911NF-10-1-0334. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the US Government. Devices were made at the UC Santa Barbara Nanofabrication Facility, a part of the NSF-funded National Nanotechnology Infrastructure Network, and at the NanoStructures Cleanroom Facility.

CONTRIBUTIONS

TCW and JYM performed the experiment and analysed the data. JD designed the experiment together with TCW, and JYM, JK, AEM and RB fabricated the sample. TCW, JYM and JMM co-wrote the manuscript. All authors contributed to the fabrication process, experimental setup and manuscript preparation.

COMPETING INTERESTS

The authors declare no conflict of interest.

REFERENCES

1. Kraus, K. *States, Effects and Operations* (Springer, 1983); URL: http://cds.cern.ch/record/98619.
2. Aharonov, Y., Albert, D. Z. & Vaidman, L. How the result of a measurement of a component of the spin of a spin-1/2 particle can turn out to be 100. *Phys. Rev. Lett.* **60**, 1351 (1988).
3. Hatridge, M. *et al.* Quantum back-action of an individual variable-strength measurement. *Science* **339**, 178–181 (2013).
4. Ruskov, R., Korotkov, A. N. & Mizel, A. Signatures of quantum behavior in single-qubit weak measurements. *Phys. Rev. Lett.* **96**, 200404 (2006).
5. Jordan, A. N., Korotkov, A. N. & Büttiker, M. Leggett-Garg inequality with a kicked quantum pump. *Phys. Rev. Lett.* **97**, 026805 (2006).
6. Williams, N. S. & Jordan, A. N. Weak values and the Leggett-Garg inequality in solid state qubits. *Phys. Rev. Lett.* **100**, 026804 (2008).
7. Palacios-Laloy, A. *et al.* Experimental violation of a Bell/'s inequality in time with weak measurement. *Nat. Phys.* **6**, 442–447 (2010).
8. Goggin, M. E. *et al.* Violation of the leggett-garg inequality with weak measurements of photons. *Proc. Natl Acad. Sci. USA* **108**, 1256–1261 (2011).
9. Dressel, J., Broadbent, C., Howell, J. & Jordan, A. Experimental violation of two-party Leggett-Garg inequalities with semiweak measurements. *Phys. Rev. Lett.* **106**, 040402 (2011).
10. Groen, J. *et al.* Partial-measurement backaction and nonclassical weak values in a superconducting circuit. *Phys. Rev. Lett.* **111**, 090506 (2013).
11. Emary, C., Lambert, N. & Nori, F. Leggett-garg inequalities. *Rep. Prog. Phys.* **77**, 016001 (2014).
12. Murch, K., Weber, S., Macklin, C. & Siddiqi, I. Observing single quantum trajectories of a superconducting quantum bit. *Nature* **502**, 211–214 (2013).
13. Weber, S. *et al.* Mapping the optimal route between two quantum states. *Nature* **511**, 570–573 (2014).
14. Bell, J. S. *et al.* On the einstein-podolsky-rosen paradox. *Physics* **1**, 195–200 (1964).
15. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880 (1969).
16. Aspect, A., Grangier, P. & Roger, G. Experimental realization of einstein-podolsky-rosen-bohm gedankenexperiment: a new violation of Bell's inequalities. *Phys. Rev. Lett.* **49**, 91 (1982).
17. Clauser, J. F. & Shimony, A. Bell's theorem. experimental tests and implications. *Rep. Prog. Phys.* **41**, 1881 (1978).

18. Dressel, J. & Korotkov, A. N. Avoiding loopholes with hybrid Bell-Leggett-Garg inequalities. *Phys. Rev. A* **89**, 012125 (2014).
19. Roch, N. *et al.* Observation of measurement-induced entanglement and quantum trajectories of remote superconducting qubits. *Phys. Rev. Lett.* **112**, 170501 (2014).
20. Leggett, A. J. & Garg, A. Quantum mechanics versus macroscopic realism: Is the flux there when nobody looks? *Phys. Rev. Lett.* **54**, 857 (1985).
21. Larsson, J.-Å. Bell's inequality and detector inefficiency. *Phys. Rev. A* **57**, 3304 (1998).
22. Ansmann, M. *et al.* Violation of bell's inequality in Josephson phase qubits. *Nature* **461**, 504–506 (2009).
23. Pearle, P. M. Hidden-variable example based upon data rejection. *Phys. Rev. D* **2**, 1418 (1970).
24. Higgins, B., Palsson, M., Xiang, G., Wiseman, H. & Pryde, G. Using weak values to experimentally determine "negative probabilities" in a two-photon state with Bell correlations. *Phys. Rev. A* **91**, 012113 (2015).
25. Fowler, A. G., Mariantoni, M., Martinis, J. M. & Cleland, A. N. Surface codes: Towards practical large-scale quantum computation. *Phys. Rev. A* **86**, 032324 (2012).
26. Kelly, J. *et al.* State preservation by repetitive error detection in a superconducting quantum circuit. *Nature* **519**, 66–69 (2015).
27. Marcovitch S., Reznik B. Testing Bell inequalities with weak measurements. Preprint at http://arxiv.org/abs/1005.3236 (2010).
28. Barends, R. *et al.* Superconducting quantum circuits at the surface code threshold for fault tolerance. *Nature* **508**, 500–503 (2014).
29. Chen, Y. *et al.* Multiplexed dispersive readout of superconducting phase qubits. *Appl. Phys. Lett.* **101**, 182601–182601 (2012).
30. Mutus, J. *et al.* Strong environmental coupling in a Josephson parametric amplifier. *Appl. Phys. Lett.* **104**, 263513 (2014).

Supplementary Information accompanies the paper on the *npj Quantum Information* website (http://www.nature.com/npjqi)

PTEP

Prog. Theor. Exp. Phys. **2015**, 021A01 (6 pages)
DOI: 10.1093/ptep/ptv005

Letter

Entanglement entropies of non-equilibrium finite-spin systems

Koichi Nakagawa*

Laboratory of Physics, School of Pharmacy and Pharmaceutical Sciences, Hoshi University, Tokyo 142-8501, Japan
*E-mail: nakagawa@hoshi.ac.jp

Received October 28, 2014; Revised December 18, 2014; Accepted December 29, 2014; Published February 6, 2015

For the purpose of clarifying a new approach to understanding quantum entanglement using thermofield dynamics (TFD), entanglement entropies of non-equilibrium finite-spin systems are examined for both traditional and extended cases. The extended entanglement entropy, $\hat{S}$, is derived, and it is found that the conditions for the maximum entangled state can be obtained through this approach. The capacity of the TFD-based method to distinguish between states in quantum systems is confirmed.

Subject Index A58

Recently, a new approach to understanding quantum entanglement using thermofield dynamics (TFD) [1–3] has been proposed in Ref. [4]. In this new treatment of quantum entanglement with TFD, extended density matrices have been formulated on the double Hilbert space (ordinary and tilde Hilbert spaces), and the entanglement states show a quantum-mechanically complicated behavior. The new TFD-based method allows the entanglement states to be easily understood, because the states in the TFD tilde space play the role of tracers of the initial states. In the new analysis, a general formulation of the extended density matrices has been constructed and applied to some simple cases. Consequently, it has been found that intrinsic quantum entanglement can be distinguished from the thermal fluctuations included in the definition of ordinary quantum entanglement at finite temperatures. Based on the analysis presented in Ref. [4], it has been argued that the new TFD-based method is applicable not only to equilibrium states but also to non-equilibrium states. However, analysis of the entanglement entropies of non-equilibrium systems was not conducted in Ref. [4] and, therefore, examination of the entanglement entropies of non-equilibrium systems with the use of TFD is of current interest. In the present communication, therefore, the "extended" entanglement entropies of non-equilibrium spin systems are intensively investigated in both the dissipative and non-dissipative cases, based upon a TFD algorithm.

Let us consider the $S = 1/2$ spin system described by the Hamiltonian

$$\mathcal{H} = -J\boldsymbol{S}_{\mathrm{A}} \cdot \boldsymbol{S}_{\mathrm{B}}, \tag{1}$$

incorporating the spin operators $\boldsymbol{S}_{\mathrm{A}} = \left(S_{\mathrm{A}}^{x}, S_{\mathrm{A}}^{y}, S_{\mathrm{A}}^{z}\right)$ and $\boldsymbol{S}_{\mathrm{B}} = \left(S_{\mathrm{B}}^{x}, S_{\mathrm{B}}^{y}, S_{\mathrm{B}}^{z}\right)$ of the subsystems A and B, respectively. The state, $|s\rangle$, of the total system is then denoted by the direct product,

$|s\rangle = |s_{\rm A}, s_{\rm B}\rangle = |s_{\rm A}\rangle|s_{\rm B}\rangle$. Using the base $\{|++\rangle, |+-\rangle, |-+\rangle, |--\rangle\}$ the matrix form of the Hamiltonian (1) is then expressed as

$$\begin{aligned}\mathcal{H} &= \sum_{s_{\rm A},s_{\rm B},s'_{\rm A},s'_{\rm B}} h_{s_{\rm A},s_{\rm B},s'_{\rm A},s'_{\rm B}} \left|s_{\rm A}, s_{\rm B}\right\rangle \left\langle s'_{\rm A}, s'_{\rm B}\right| \\ &= -\frac{J}{4}\left(|++\rangle\langle++| + |--\rangle\langle--|\right) + \frac{J}{4}\left(|+-\rangle\langle+-| + |-+\rangle\langle-+|\right) \\ &\quad - \frac{J}{2}\left(|+-\rangle\langle-+| + |-+\rangle\langle+-|\right). \end{aligned} \tag{2}$$

For the equilibrium states in terms of the Hamiltonian expressed in Eq. (2) , the ordinary density matrix, $\rho_{\rm eq}$, of this system can be obtained as

$$\begin{aligned}\rho_{\rm eq} := \frac{e^{-\beta\mathcal{H}}}{Z(\beta)} = \frac{e^{-\beta J/4}}{Z(\beta)}\Bigg(& e^{\beta J/2}\left(|++\rangle\langle++| + |--\rangle\langle--|\right) \\ &+ \cosh\frac{\beta J}{2}\left(|+-\rangle\langle+-| + |-+\rangle\langle-+|\right) \\ &+ \sinh\frac{\beta J}{2}\left(|-+\rangle\langle+-| + |+-\rangle\langle-+|\right)\Bigg), \end{aligned} \tag{3}$$

where β is the inverse temperature and the partition function, $Z(\beta)$, is defined as

$$Z(\beta) := {\rm Tr} e^{-\beta\mathcal{H}} = 2e^{-\beta J/4}\left(e^{\beta J/2} + \cosh\frac{\beta J}{2}\right). \tag{4}$$

Let us turn our attention to a non-equilibrium system with dissipation, which is described by the Hamiltonian of Eq. (2). The time dependence of the ordinary density matrix, $\rho(t)$, of this system is given by the dissipative von Neumann equation [5–7], where

$$i\hbar\frac{\partial}{\partial t}\rho(t) = \left[\mathcal{H}, \rho(t)\right] - \epsilon\left(\rho(t) - \rho_{\rm eq}\right), \tag{5}$$

with ϵ being a dissipation parameter. The solution of Eq. (5) is expressed as

$$\rho(t) = e^{-\epsilon t}U^{\dagger}(t)\rho_0 U(t) + \left(1 - e^{-\epsilon t}\right)\rho_{\rm eq}, \tag{6}$$

for any initial density matrix, ρ_0, where the unitary operator, $U(t)$, denotes

$$\begin{aligned}U(t) := e^{i\mathcal{H}t/\hbar} = e^{i\omega t/4}\Bigg(& e^{-i\omega t/2}\left(|++\rangle\langle++| + |--\rangle\langle--|\right) \\ &+ \cos\frac{\omega t}{2}\left(|+-\rangle\langle+-| + |-+\rangle\langle-+|\right) \\ &- i\sin\frac{\omega t}{2}\left(|-+\rangle\langle+-| + |+-\rangle\langle-+|\right)\Bigg), \end{aligned} \tag{7}$$

and $\omega := J/\hbar$. Because the explicit expression of $\rho(t)$ in Eq. (6) is complicated for any initial condition, hereafter, let us confine ourselves to the initial condition $\rho_0 = |+-\rangle\langle+-|$. Inserting Eqs. (3) and (7), along with the initial condition, into Eq. (6), we then obtain

$$\begin{aligned}\rho(t) = \frac{e^{-\epsilon t}}{2}\Bigg(&\frac{2\left(e^{\epsilon t}-1\right)}{3+e^{-\beta\omega\hbar}}\left(|++\rangle\langle++|+|--\rangle\langle--|\right)\\ &+\frac{\cos\omega t+e^{\epsilon t}+e^{\beta\omega\hbar}\left(3\cos\omega t+e^{\epsilon t}+2\right)}{1+3e^{\beta\omega\hbar}}|+-\rangle\langle+-|\\ &+\frac{-\cos\omega t+e^{\epsilon t}+e^{\beta\omega\hbar}\left(-3\cos\omega t+e^{\epsilon t}+2\right)}{1+3e^{\beta\omega\hbar}}|-+\rangle\langle-+|\\ &+\left(\frac{\left(e^{\epsilon t}-1\right)\left(-1+e^{\beta\omega\hbar}\right)}{1+3e^{\beta\omega\hbar}}-i\sin\omega t\right)|+-\rangle\langle-+|\\ &+\left(\frac{\left(e^{\epsilon t}-1\right)\left(-1+e^{\beta\omega\hbar}\right)}{1+3e^{\beta\omega\hbar}}+i\sin\omega t\right)|-+\rangle\langle+-|\Bigg).\end{aligned} \tag{8}$$

The ordinary entanglement entropy, S, is defined by

$$S := -k_{\mathrm{B}}\mathrm{Tr}_{\mathrm{A}}\left[\rho_{\mathrm{A}}\log\rho_{\mathrm{A}}\right], \tag{9}$$

with $\rho_{\mathrm{A}} := \mathrm{Tr}_{\mathrm{B}}\rho(t)$, where Tr_{A} and Tr_{B} represent traces over the variables of subsystems A and B, respectively. The insertion of Eq. (8) into Eq. (9) yields

$$S = -k_{\mathrm{B}}\left(\frac{1+e^{-\epsilon t}\cos\omega t}{2}\log\frac{1+e^{-\epsilon t}\cos\omega t}{2}+\frac{1-e^{-\epsilon t}\cos\omega t}{2}\log\frac{1-e^{-\epsilon t}\cos\omega t}{2}\right). \tag{10}$$

It is also possible to argue that S in Eq. (10) is directly proportional to an entanglement, $E(C)$, which is a function of the "concurrence", $C := \sqrt{1-e^{-2\epsilon t}\cos^2\omega t}$ [8]. The time dependence of S and C is displayed in Fig. 1 (in units of $k_{\mathrm{B}} = 1$). In the dissipative system, S and C converge to the constants $k_{\mathrm{B}}\log 2$ and 1, respectively, at $t\to\infty$, so it is reasonable to think that S and C include not only the contribution of the quantum fluctuation, but also the contribution of the classical and thermal fluctuations. However, this fact is not manifest in the above expressions of S and C.

We are now in a position to investigate the extended density matrix, $\hat{\rho}$, in the TFD double Hilbert space. Note that $\hat{\rho}$ has been defined in Ref. [4] as follows:

$$\hat{\rho} := |\Psi\rangle\langle\Psi|,\ \ |\Psi\rangle := \rho(t)^{1/2}\sum_s |s,\tilde{s}\rangle = \rho(t)^{1/2}\sum_s |s\rangle|\tilde{s}\rangle, \tag{11}$$

using the ordinary density matrix, $\rho(t)$, in Eq. (6), where $\{|s\rangle\}$ is the orthogonal complete set in the original Hilbert space and $\{|\tilde{s}\rangle\}$ is the same set in the tilde Hilbert space of the TFD [9,10]. If entanglement subsystems A and B are being examined, each of the $|s\rangle$ and $|\tilde{s}\rangle$ states are represented as the direct products $|s_{\mathrm{A}}, s_{\mathrm{B}}\rangle = |s_{\mathrm{A}}\rangle|s_{\mathrm{B}}\rangle$ and $|\tilde{s}_{\mathrm{A}}, \tilde{s}_{\mathrm{B}}\rangle = |\tilde{s}_{\mathrm{A}}\rangle|\tilde{s}_{\mathrm{B}}\rangle$, respectively. We are then led to the renormalized extended density matrix, $\hat{\rho}_{\mathrm{A}}$, as

$$\begin{aligned}\hat{\rho}_{\mathrm{A}} &:= \mathrm{Tr}_{\mathrm{B}}\hat{\rho} := \sum_{s_{\mathrm{B}},\tilde{s}'_{\mathrm{B}}}\left\langle s_{\mathrm{B}},\tilde{s}'_{\mathrm{B}}|\hat{\rho}|s_{\mathrm{B}},\tilde{s}'_{\mathrm{B}}\right\rangle\\ &= b_{\mathrm{d1}}|+\rangle\langle+||\tilde{+}\rangle\langle\tilde{+}|+b_{\mathrm{d2}}|-\rangle\langle-||\tilde{-}\rangle\langle\tilde{-}|+b_{\mathrm{cf}}\left(|+\rangle\langle-||\tilde{+}\rangle\langle\tilde{-}|+|-\rangle\langle+||\tilde{-}\rangle\langle\tilde{+}|\right)\\ &\quad+b_{\mathrm{qe}}\left(|+\rangle\langle+||\tilde{-}\rangle\langle\tilde{-}|+|-\rangle\langle-||\tilde{+}\rangle\langle\tilde{+}|\right),\end{aligned} \tag{12}$$

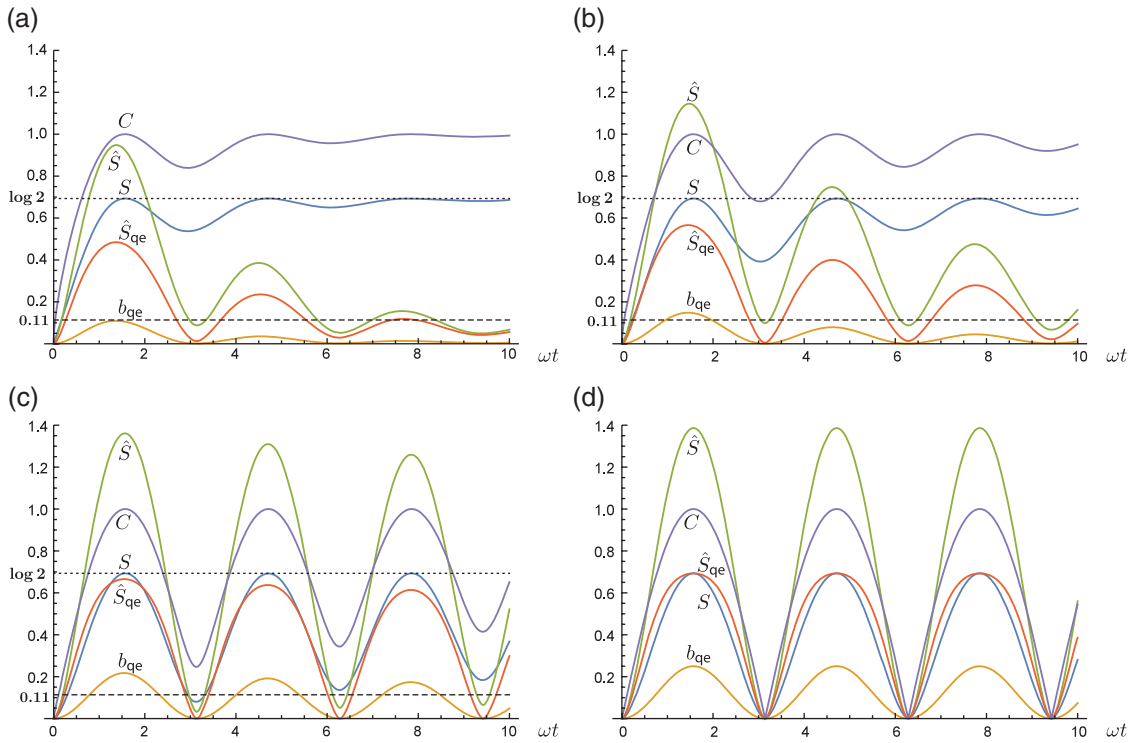


Fig. 1. Time dependence of entropies C, S, $\hat{S}$, and $\hat{S}_{\rm qe}$, along with parameter $b_{\rm qe}$, in dissipative and non-dissipative systems with scaled temperature, $T/J = 0.7$. Parts (a), (b), and (c) show cases with a scaled dissipation rate of $\epsilon/\omega = 0.2$, 0.1, and 0.01, respectively. The dotted and dashed lines in parts (a), (b), and (c) represent the asymptotes of the S and $\hat{S}$ curves, respectively. Part (d) is the non-dissipation case ($\epsilon = 0$).

where the matrix elements $b_{\rm d1}$, $b_{\rm d2}$, $b_{\rm cf}$, and $b_{\rm qe}$ are respectively obtained as analytic functions of t, β, ϵ, and ω, and correspond to the two diagonal components (d1 and d2), the classical fluctuations (cf), and the quantum entanglements (qe) of $\hat{\rho}_{\rm A}$, respectively. The parameter $b_{\rm qe}$ in Eq. (12) expresses the quantum entanglement effect. This quantum fluctuation plays a crucial role in various quantum systems, and it has been used as an order parameter of 2D quantum systems [11,12]. The time dependences of the parameter $b_{\rm qe}$ in several cases are shown in Fig. 1. As can be seen from Eq. (12), only the intrinsic quantum entanglement is extracted clearly in the TFD formulation. In particular, it can be understood that the entangled state of the system emerges through a single product, such as $|+\rangle\langle+||\tilde{-}\rangle\langle\tilde{-}|$, in $\hat{\rho}_{\rm A}$.

The "extended" entanglement entropy is defined as

$$\hat{S} := -k_{\rm B}{\rm Tr}_{\rm A}\left[\hat{\rho}_{\rm A}\log\hat{\rho}_{\rm A}\right], \tag{13}$$

using the renormalized $\hat{\rho}_{\rm A}$ in Eq. (12) [4]. The insertion of Eq. (12) into Eq. (13) and subsequent simplification eventually yield

$$\hat{S} = \hat{S}_{\rm cl} + \hat{S}_{\rm qe}, \tag{14}$$

where

$$\hat{S}_{\rm cl} := -k_{\rm B}\left(\sqrt{4b_{\rm cf}^2 + \left(b_{\rm d1} - b_{\rm d2}\right)^2}\,{\rm arccoth}\,\frac{b_{\rm d1} + b_{\rm d2}}{\sqrt{4b_{\rm cf}^2 + \left(b_{\rm d1} - b_{\rm d2}\right)^2}} + \frac{b_{\rm d1} + b_{\rm d2}}{2}\log\left(b_{\rm d1}b_{\rm d2} - b_{\rm cf}^2\right)\right), \tag{15}$$

and

$$\hat{S}_{\rm qe} := -2k_{\rm B} b_{\rm qe} \log b_{\rm qe}, \tag{16}$$

respectively. In Eqs. (14), (15), and (16), the expressions of $\hat{S}$, the classical and thermal fluctuation parts, $\hat{S}_{\rm cl}$, and the quantum entanglement part, $\hat{S}_{\rm qe}$, also incorporate analytic functions of t, β, ϵ, and ω, respectively; however, the full calculation is quite tedious. So, we show the numerical behavior of C, S, $\hat{S}$, $\hat{S}_{\rm qe}$, and $b_{\rm qe}$ for a few cases in Figs. 1(a)–(c) (in units of $k_{\rm B} = 1$). As can be seen from these figures, at $t \to \infty$, $\hat{S}$ converges to the value $0.113\,51\cdots$, and both $\hat{S}_{\rm qe}$ and $b_{\rm qe}$ vanish, respectively. As a consequence, the traditional entanglement entropy, S, becomes larger than the extended entanglement entropies, $\hat{S}$ and $\hat{S}_{\rm qe}$, at $t \to \infty$. $\hat{S}_{\rm qe}$ is then smaller than S when ϵ is relatively larger. As ϵ becomes smaller, $\hat{S}_{\rm qe}$ becomes compatible with S and, at $\epsilon = 0$, $\hat{S}_{\rm qe} \lesssim S$. These results suggest that the quantum entanglement is enhanced as the dissipation becomes weaker.

For non-dissipative systems, $\hat{S}$ in Eq. (14) and $\hat{S}_{\rm qe}$ in Eq. (16) reduce to

$$\hat{S} = -k_{\rm B}\left(\cos^4\frac{\omega t}{2}\cdot\log\left(\cos^4\frac{\omega t}{2}\right) + \sin^4\frac{\omega t}{2}\cdot\log\left(\sin^4\frac{\omega t}{2}\right) + \frac{1}{2}\sin^2\omega t\cdot\log\left(\frac{\sin^2\omega t}{4}\right)\right), \tag{17}$$

and

$$\hat{S}_{\rm qe} = -\frac{k_{\rm B}}{2}\sin^2\omega t\cdot\log\left(\frac{\sin^2\omega t}{4}\right), \tag{18}$$

respectively, at $\epsilon = 0$. The time dependence of $\hat{S}$ and $\hat{S}_{\rm qe}$ at $\epsilon = 0$ is shown in Fig. 1(d) (in units of $k_{\rm B} = 1$). It is apparent in this figure that all the curves (C, S, $\hat{S}$, $\hat{S}_{\rm qe}$, and $b_{\rm qe}$) showing the entanglement have the same phase; however, their amplitudes differ. Specifically, $\hat{S}$ is larger than S and $\hat{S}_{\rm qe} \approx S$ at $\epsilon = 0$, a result that differs from that of Ref. [4] and that can be seen in Eqs. (17) and (18). It appears that a mistake was made in Ref. [4] in counting the non-zero eigenvalues of $\hat{\rho}_{\rm A}$.

In this communication, we have examined the extended entanglement entropies of non-equilibrium spin systems in both the dissipative and non-dissipative cases, based upon the TFD formulation. In the dissipative case in particular, the extended entanglement entropy is derived using the extended density matrix and is proven to separate into the classical and thermal fluctuation parts and the quantum entanglement part. These quantities are compared to the traditional entanglement entropy, the concurrence, and $b_{\rm qe}$ in $\hat{\rho}_{\rm A}$. These results are summarized in Fig. 1 and show that the conditions yielding the maximum entangled state can be obtained using these five quantities.

We have clearly indicated that, in the TFD formulation, the extended quantum entanglement entropy part and the parameter $b_{\rm qe}$ are recognized as effective quantities for measurement of the quantum entanglement. It is apparent that the new TFD-based method enables us to clearly distinguish between the various states of quantum systems.

References

[1] U. Fano, Mod. Phys. A, **42**, 74 (1957).
[2] I. Prigogine, C. George, F. Henin, and L. Rosenfeld, Chem. Scripta, **4**, 5 (1973).
[3] Y. Takahashi and H. Umezawa, Collect. Phenom., **2**, 55 (1975).
[4] Y. Hashizume and M. Suzuki, Physica A, **392**, 3518 (2013).

[5] M. Suzuki, In *Quantum bio-informatics II: from quantum information to bio-informatics, Tokyo University of Science, Japan, 12–16 March 2008 (QP-PQ: Quantum probability and white noise analysis; v. 24)*, eds. L. Accardi, W. Freundberg, and M. Ohya (World Scientific, Singapore, 2008).
[6] M. Suzuki, Physica A, **390**, 1904 (2011).
[7] M. Suzuki, Physica A, **391**, 1074 (2012).
[8] W. K. Wootters, Phys. Rev. Lett., **80**, 2245 (1998).
[9] M. Suzuki, J. Phys. Soc. Jpn., **54**, 4483 (1985).
[10] M. Suzuki, *Statistical Mechanics* (Iwanami, Tokyo, 2000).
[11] J.-M. Stephan, S. Furukawa, G. Misguich, and V. Pasquier, Phys. Rev. B, **80**, 184421 (2009).
[12] S. Tanaka, R. Tamura, and H. Katsura, Phys. Rev. A, **86**, 032326 (2012).

CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 074 **December 31, 2017**

Exact Equivariance, Disentanglement and Invariance of Transformations

by

Qianli Liao and Tomaso Poggio
Center for Brains, Minds, and Machines, McGovern Institute for Brain Research,
Massachusetts Institute of Technology, Cambridge, MA, 02139.

Abstract: Invariance, equivariance and disentanglement of transformations are important topics in the field of representation learning. Previous models like Variational Autoencoder [1] and Generative Adversarial Networks [2] attempted to learn disentangled representations from data with different levels of successes. Convolutional Neural Networks are *approximately* equivariant and invariant (if pooling is performed) to input translations. In this report, we argue that the recently proposed Object-Oriented Learning framework [3] offers a new solution to the problem of Equivariance, Invariance and Disentanglement: it systematically factors out common transformations like translation and rotation in inputs and achieves **"exact equivariance"** to these transformations — that is, when the input is translated and/or rotated by some amount, the output and all intermediate representations of the network are also translated and rotated by **exactly the same amount**. The transformations are **"exactly disentangled"** in the sense that the translations and rotations can be read out directly from a few known variables of the system without any approximation. Invariance can be achieved by reading other variables that are known not to be affected by the transformations. No learning is needed to achieve these properties. Exact equivariance and disentanglement are useful properties that augment the expressive power of neural networks. We believe it will enable new applications including but not limited to precise visual localization of objects and measuring of motion and angles.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF - 1231216.

Contents

1 Introduction **3**

1.1 Common Approaches to Invariance, Equivariance and Disentanglement 3

1.2 Object-Oriented Learning: A New Solution to Invariance, Equivariance and Disentanglement 3

1.3 Equivariance and Disentanglement with Arbitrarily Fine Precision 4

2 Model **4**

2.1 Voting Layer 5

2.2 Binding Layer 5

3 Experiments **5**

3.1 Examples of Exact Equivariance 5

3.2 Exact Equivariance in 3D 5

3.3 Invariant Recognition of 2D Patterns 9

1 Introduction

We first discuss our interpretation of the following three concepts: Invariance, Equivariance and Disentanglement. **Invariance** is a property that the representation does not change under some transformations of the input. It has been studied extensively by I-theory and related work [4, 5, 6].

Equivariance is a property that the representation changes when input transforms (see [7] for related discussion). Note that equivariance under this general definition is easy to achieve — even random projections of the input or the input itself are equivariant. When we use the term equivariance, it stresses the fact that information of the transformations is retained/encoded by the network, instead of being discarded, contrasting the case of invariant representations.

Disentanglement, on the other hand, is highly related to equivariance but being more specific. It is a property that the information of some transformation of the input is not only encoded in the network/system, but also in only a few variables instead of being spreaded out in a distributed fashion. In other words, transforming the input should change the representation extracted by the network, but only affecting a few (preferably known) dimensions, leaving the other dimensions completely unchanged.

Disentanglement is a perfect balance of invariance and equivariance: for any transformation, the representation provides both invariant and equivariant code, but in separate (known) channels/dimensions. This allows for clear, versatile and interpretable manipulation of information throughout the system.

Notice that existing invariance scheme such as in convolutional networks or the more general ones described in i-theory can be made equivariant, invariant and disentagled by performing a read-out pooling operation at the very last layer.

1.1 Common Approaches to Invariance, Equivariance and Disentanglement

Translation is the most common transformation in a wide variaty of modalities. The most popular solution to invariance and equivariance of translation is the well-known convolutional neural networks (ConvNets) [8, 9, 10]. If spatial pooling is performed, the representation becomes invariant to local or global translation of the input.

Although ConvNets can be equivariant to translation, there is significant amount of quantization error as the signals travel from bottom to top of the network (even without local pooling). In higher layers, the spatial locations of inputs are not encoded in a precise fashion — the spatial uncertainty accumulates as the number of layers increases. Local and global poolings, if added to ConvNet, discard translation information in a unrecoverable fashion, making this problem even worse.

There has been a large body of literature (to be discussed more later) focusing on learning disentangled representations from data using Variational Autoencoder [1] and Generative Adversarial Networks [2]. They tend to show that some dimensions in high layers of the network encodes common transformations like translation, rotation and some more general transformations in a mixed fashion on digits (MNIST) [1] and recently on domain-specific datasets [11, 12]. As most of these models adopt ConvNets, they suffer from the same issue of not being precise (e.g., loss of translation information in higher layers). Hinton's Capsule model [13] also learns some level of disentangled representations from data. The limitation of above approaches are that they are very data dependent and have not been generalized well to broad categories of real-world objects (e.g., ImageNet). And it is not clear how to systematically deal with several transformations without interference.

1.2 Object-Oriented Learning: A New Solution to Invariance, Equivariance and Disentanglement

Object-Oriented Learning [3] provides a new approach to the problem of invariance, equivariance and disentanglement. Instead of encoding transformations into distributed code as that in the traditional "feature-oriented" learning, we

adopt a basic representational atom "object/symbol" that packages transformation parameters as fields/properties (conceptually analogous to a class in object-oriented programming).

In our current system, each object has several fields/properties: its location represented by x and y coordinates (and z if extended into 3D), its pose of rotation represented by an angle (or quaternion in the case of 3D), its scale represented by a scaling factor and finally its signature, which is a vector that encodes information that can identify the object but invariant to its transformations (e.g., translation, rotation and scaling).

Each field of an object represents a disentangled transformation parameter. When an object is predicted, its transformation parameters are also predicted. The transformation parameters and signature of objects can be trained end-to-end on any task.

1.3 Equivariance and Disentanglement with Arbitrarily Fine Precision

The main advantage of our treatment of equivariance and disentanglement is that in our framework they can be made **"arbitrarily precise"**. Unlike ConvNets, which are restricted by its grid representation and inability to represent rotations and scalings, our framework encodes transformation parameters as fields of the objects, with arbitrary precision (e.g., floating point numbers in our experiments).

Our framework offers a new level of expressive power: whenever the input is transformed (e.g., translated, rotated or scaled [1]) by some amount T_g, the representation in every intermediate layer of our network reflects **exactly the same amount** of transformation T_g in the objects' fields.

Property 1: Exact Equivariance: Let we use x to denote the input to the system (or to any intermediate layers of it). Let $F()$ be the function of arbitrary number of "object-oriented layers", $T_g()$ be a transformation function parameterized by transformation parameter g. We have:

$$T_g(F(x)) \equiv F(T_g(x)) \tag{1}$$

Essentially, the transformation commutes with arbitrary number of layers of processing of our object-oriented layers. In our current implementation, T() currently supports a combination of translations and rotations (in both 2D and 3D).

Property 2: Exact Disentanglement: there exist a read-out function R() that can always extract transformation parameters by comparing the **high level representations** before and after a transformation of the **input**:

$$R(F(T_g(x)), F(x)) \equiv g \tag{2}$$

Note that R() should be as simple as possible (e.g., reading only several known dimensions of the high-level representation) and ideally be non-parametric since if substantial learning is required to read transformation parameters, it can hardly be called a "disentangled representation".

In above definitions, our models guarantees $\equiv$, that is why we name them "exact" equivariance and disentanglement. In previous models (like ConvNets), $\equiv$ may become $\approx$, thus becoming not exact.

2 Model

We describe the mechanisms that are needed to achieve Property 1 and 2.

[1] Scaling is currently treated in the system as several discrete scales. Continuous scaling is being implemented.

2.1 Voting Layer

The voting layer is described very briefly in [3] and in a more general and principled way in [14]. The main conclusion from [14] is the following claim:

Proposition 1: Commutative Property of Voting Layer

Assume the input x are a list of objects (instead of pixels), the voting layer $V()$ commutes with the transformation $T()$. That is:

$$T(V(x)) = V(T(x)) \tag{3}$$

Note that T(x) means that the same transformation is applied to every element of the list of objects x.

2.2 Binding Layer

The binding layer is different from that in [3]. The main idea is to select a subset of objects as representatives of local clusters of using a suppression mechanism, and these representatives will bind neighbor objects (and itself) to form a new object. It has two meta parameters: binding radius r_b and suppression radius r_s. There are different binding procedures within this class that can satisfy the exact equivariance property, but the the particular binding procedure we implemented is: 1. every object counts the number of neighbors within radius r_b in euclidean space (2D or 3D) and compute the sum of distances from neighbors to itself. 2. Then every object looks at neighbors within radius r_s to see if it is the object with most neighbors (if there is a tie, check if it has the lowest distances). If so, it becomes a representative. 3. Each representative averages all objects within radius r_s. By averaging, we refer to both averaging of signatures and properties To be precise, in our implementation, we performed sum of signatures and average of properties. Average of signatures should behave qualitatively similar, but we did not try it.

Finally, with this binding procedure, we have the additional claim:

Proposition 2: Commutative Property of Binding Layer Assume the input x are a list of objects (instead of pixels), the voting layer $B()$ commutes with the transformation $T()$. That is:

$$B(V(x)) = B(T(x)) \tag{4}$$

Note that T(x) means that the same transformation is applied to every element of the list of objects x.

Both binding and voting layers commute with the transformations (translation and rotations).

3 Experiments

3.1 Examples of Exact Equivariance

We show several examples of exact equivariance on some simple data in Figure 1, 2 and 3.

3.2 Exact Equivariance in 3D

See [14] for extension of this model in 3D We have also checked

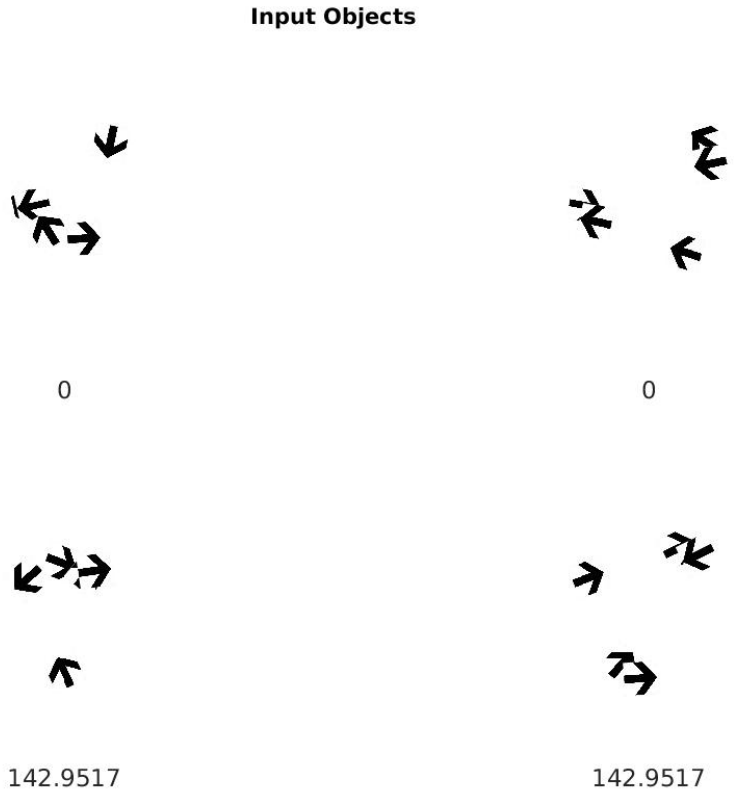


Figure 1: **Exact Equivariance Visualization:** Input objects to our network. First row shows two patterns (each pattern is a collection of objects, illustrated by arrows). Second row shows the two patterns rotated by some angles (indicated below the image). The rotated images look slightly different because the objects are drawn in some random orders — the overlay orders are different.

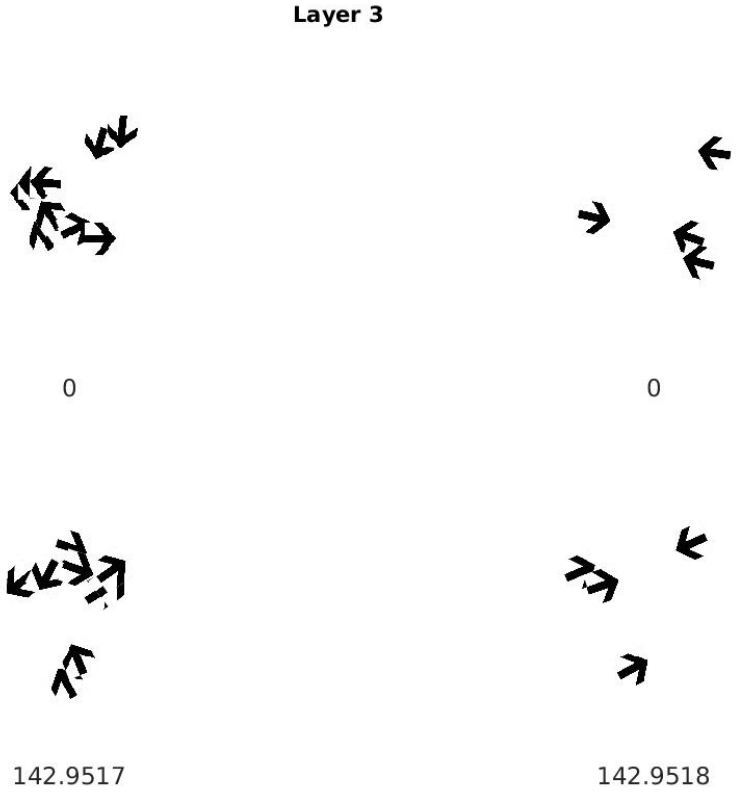


Figure 2: **Exact Equivariance Visualization:** Activation of the 3rd layer of our network. Recall that all the intermediate activations of our network are also objects. Each sub-figure corresponds to the 3rd layer activations induced by corresponding inputs in Figure 1. The numbers below the sub figures are the average of rotation angles of the objects (using first row objects as frame of reference — that is why first row values are 0s). We can see that the 3rd layer activations are exactly equivariant to rotations of the input in Figure 1. The rotated images might look slightly different because the objects are drawn in some random orders — the overlay orders are different.

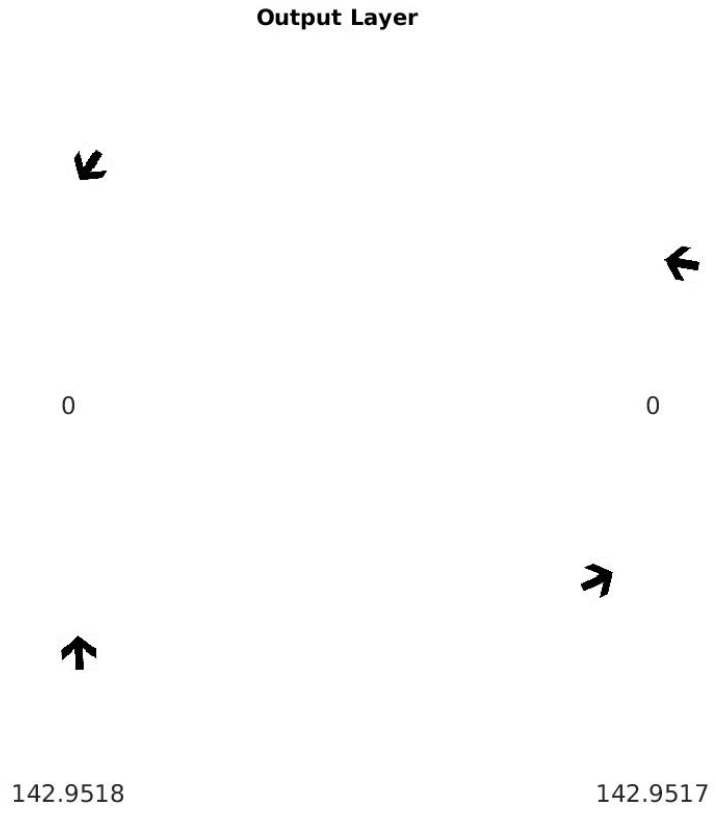


Figure 3: **Exact Equivariance Visualization:** Activation of the 6th layer of our network. Recall that all the intermediate activations of our network are also objects. Each sub-figure corresponds to the 6th layer activations induced by corresponding inputs in Figure 1. The numbers below the sub figures are the average of rotation angles of the objects (using first row objects as frame of reference — that is why first row values are 0s). We can see that the 6th layer activations are exactly equivariant to rotations of the input in Figure 1.

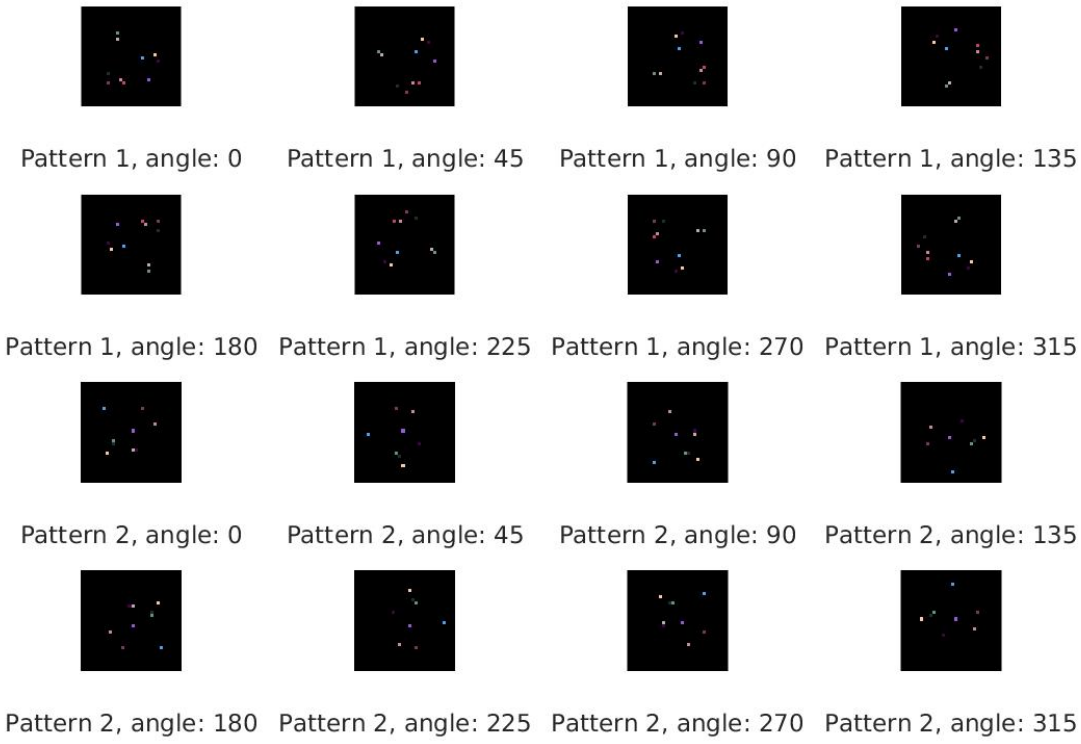


Figure 4: Visualization of 2D Patterns used in our experiments

3.3 Invariant Recognition of 2D Patterns

We train and test our model and ConvNet (the same one used in [3]) on recognizing 2D patterns. There are 10 "objects" (i.e., 10 dots in each pattern) scattered in 10 different ways (i.e., 10-way classification with chance performance being 90% error). See Figure 4 for visualization. The inputs to our model are objects. Each object has a signature (3-dimensional, RGB value), a position (x and y) and a pose (scalar angle). To compare with ConvNet, we embed the objects' signature and pose into corresponding positions of images. Since positions in images can only take integer values, we add a uniform noise in [-1,1] to the positions of the objects (different for every minibatch) and rounded them to integers — this input preprocessing procedure is done for both our model and ConvNet so that both models take in the same inputs.

The results are shown in Figure 5 and 6. Our model generalizes perfectly to novel rotations that the model has not seen before.

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[3] Q. Liao and T. Poggio, "Object-oriented deep learning," tech. rep., Center for Brains, Minds and Machines (CBMM), https://dspace.mit.edu/handle/1721.1/112103, October, 2017.

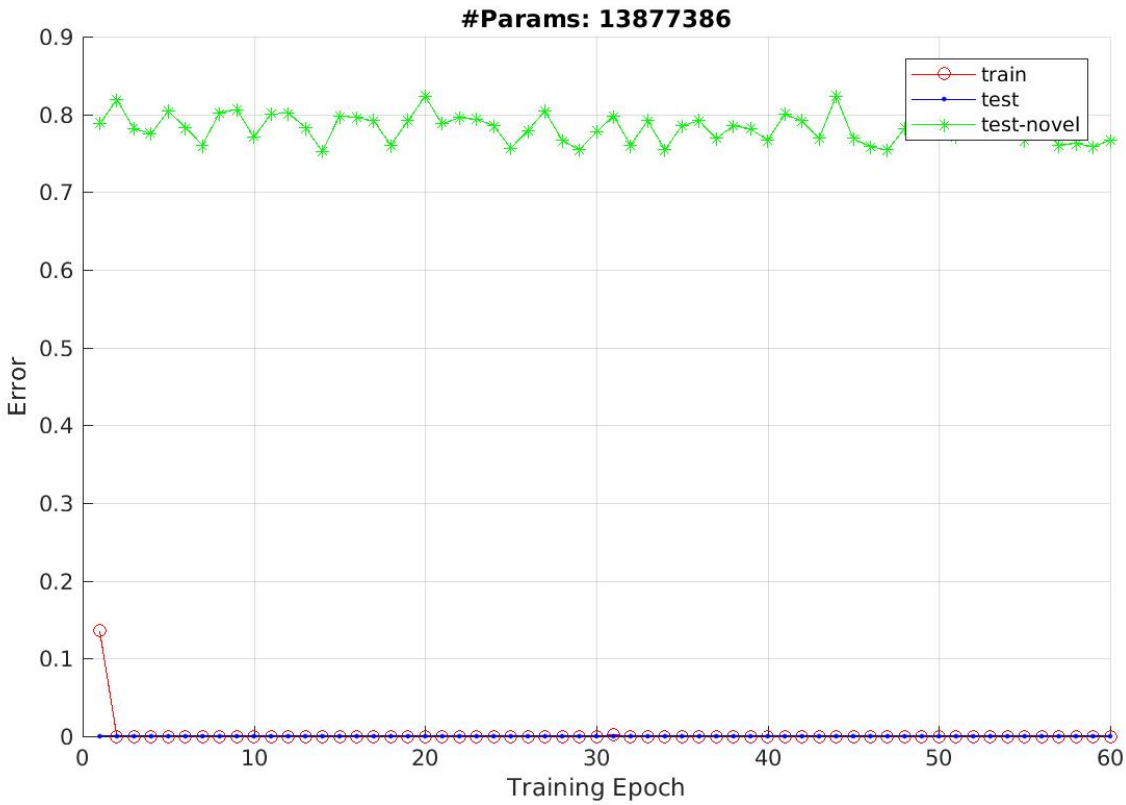


Figure 5: Performance of ConvNet (ResNet). It converges to 0 error quickly on training, but does not generalize at all to novel rotations.

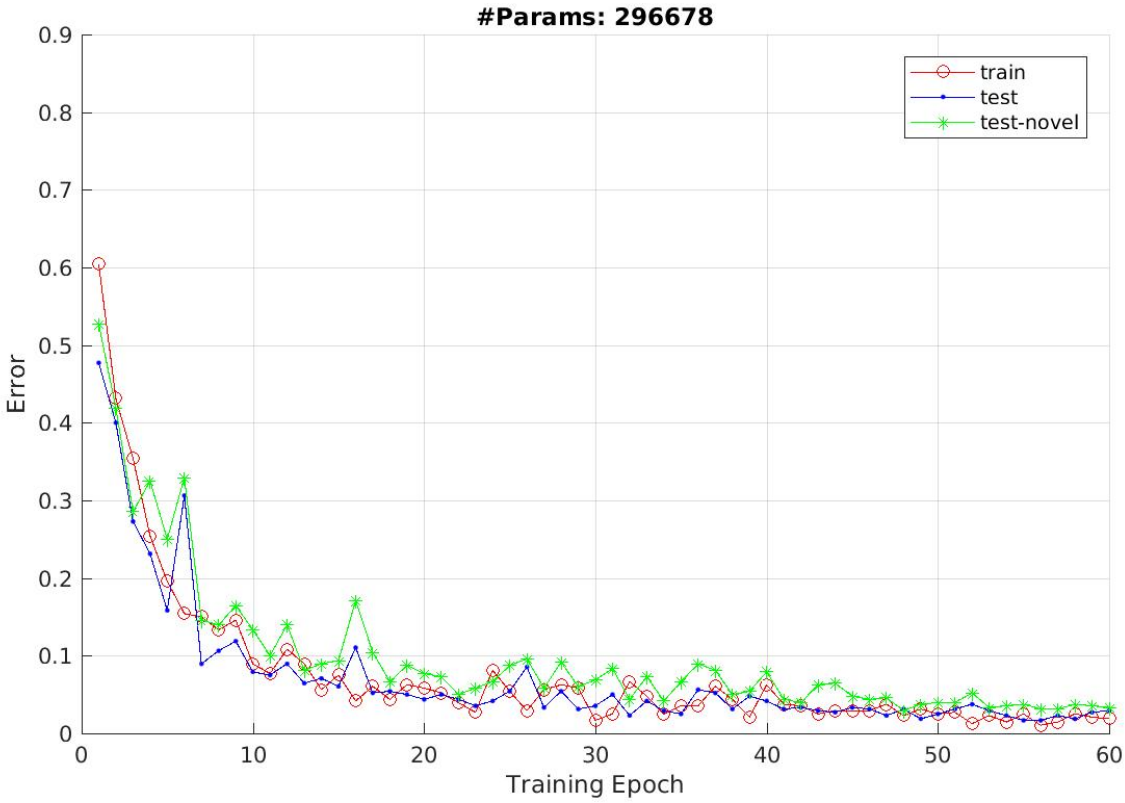


Figure 6: Performance of our model. Thanks to the exact equivariance and disentanglement property, there is no difference between the performance of training (rotation=0) test (rotation=0) and test-novel (360 degrees of rotations).

[4] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations in hierarchical architectures," *arXiv preprint arXiv:1311.4158*, 2013.

[5] Q. Liao, J. Z. Leibo, and T. Poggio, "Learning invariant representations and applications to face verification," in *Advances in Neural Information Processing Systems*, pp. 3057–3065, 2013.

[6] J. Z. Leibo, Q. Liao, F. Anselmi, and T. Poggio, "The invariance hypothesis implies domain-specific regions in visual cortex," *bioRxiv, 10.1101/004473*, 2014.

[7] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

[8] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, Apr. 1980.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, November 1998.

[10] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[11] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in Neural Information Processing Systems*, pp. 2539–2547, 2015.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[13] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *International Conference on Artificial Neural Networks*, pp. 44–51, Springer, 2011.

[14] Q. Liao and T. Poggio, "3d object-oriented learning: An end-to-end transformation-disentangled 3d representation," tech. rep., Center for Brains, Minds and Machines (CBMM), 2017.

Exact Equivariance, Disentanglement and Invariance of Transformations

Author(s)
Liao, Qianli; Poggio, Tomaso

Abstract
Invariance, equivariance and disentanglement of transformations are important topics in the field of representation learning. Previous models like Variational Autoencoder [1] and Generative Adversarial Networks [2] attempted to learn disentangled representations from data with different levels of successes. Convolutional Neural Networks are approximately equivariant and invariant (if pooling is performed) to input translations. In this report, we argue that the recently proposed Object-Oriented Learning framework [3] offers a new solution to the problem of Equivariance, Invariance and Disentanglement: it systematically factors out common transformations like translation and rotation in inputs and achieves "exact equivariance" to these transformations — that is, when the input is translated and/or rotated by some amount, the output and all intermediate representations of the network are also translated and rotated by exactly the same amount. The transformations are "exactly disentangled" in the sense that the translations and rotations can be read out directly from a few known variables of the system without any approximation. Invariance can be achieved by reading other variables that are known not to be affected by the transformations. No learning is needed to achieve these properties. Exact equivariance and disentanglement are useful properties that augment the expressive power of neural networks. We believe it will enable new applications including but not limited to precise visual localization of objects and measuring of motion and angles.

Date issued
2017-12-31

URI
http://hdl.handle.net/1721.1/113001

Series/Report no.
CBMM Memo Series;074

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 1, 2021, accepted July 25, 2021, date of publication July 30, 2021, date of current version August 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3101229

Disentangled Representation Learning in Real-World Image Datasets via Image Segmentation Prior

NAO NAKAGAWA[1], (Graduate Student Member, IEEE), REN TOGO[2], (Member, IEEE), TAKAHIRO OGAWA[3], (Senior Member, IEEE), AND MIKI HASEYAMA[3], (Senior Member, IEEE)

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[2]Education and Research Center for Mathematical and Data Science, Hokkaido University, Sapporo 060-0812, Japan
[3]Faculty of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Nao Nakagawa (nakagawa@lmd.ist.hokudai.ac.jp)

This work was partly supported by JSPS KAKENHI Grant Number JP21H03456.

ABSTRACT We propose a novel method that can learn easy-to-interpret latent representations in real-world image datasets using a VAE-based model by splitting an image into several disjoint regions. Our method performs object-wise disentanglement by exploiting image segmentation and alpha compositing. With remarkable results obtained by unsupervised disentanglement methods for toy datasets, recent studies have tackled challenging disentanglement for real-world image datasets. However, these methods involve deviations from the standard VAE architecture, which has favorable disentanglement properties. Thus, for disentanglement in images of real-world image datasets with preservation of the VAE backbone, we designed an encoder and a decoder that embed an image into disjoint sets of latent variables corresponding to objects. The encoder includes a pre-trained image segmentation network, which allows our model to focus only on representation learning while adopting image segmentation as an inductive bias. Evaluations using real-world image datasets, CelebA and Stanford Cars, showed that our method achieves improved disentanglement and transferability.

INDEX TERMS Alpha blend, disentanglement, image segmentation, real-world image, representation learning.

I. INTRODUCTION

Unsupervised representation learning in image datasets is one of the main challenges in the field of computer vision [1]–[3], and it can support downstream tasks such as transfer learning and reinforcement learning [1], [4]. A popular framework for unsupervised representation learning is a deep generative model, which aims to generate high-dimensional images from low-dimensional latent variables [1], [3]–[6]. Disentangled representation learning (DRL) aims at separating the representation of latent variables into disjoint parts corresponding to semantically meaningful features [1], [2], [4], [6], [7]. Disentangled representations can be beneficial for various tasks of computer vision such as controllable image generation [8]–[11], person identification [12], [13] and robust adversarial training [14].

Deep generative models based on a variational autoencoder (VAE) [5] were used in previous studies on DRL. A VAE model has an encoder that infers the posterior distribution of latent variables [5]. Hence, in a VAE-based model, independence constraints can be explicitly imposed on the latent variables by using the output posterior distribution. Several recently proposed VAE-based DRL methods disentangle the latent representation by imposing independence constraints in the loss function [6].

Various DRL methods based on the information bottleneck have been proposed for disentangling the latent variables [4], [6], [15]–[21]. In many DRL methods used in previous studies, attempts were made to discover factors of variation without any prior information or supervision. Although remarkable results were obtained by these unsupervised DRL

The associate editor coordinating the review of this manuscript and approving it for publication was Khursheed Aurangzeb.

methods using toy datasets such as dSprites [22] and 3D Shapes [23], there is no guarantee that each latent variable corresponds to a single semantically meaningful factor of variation without any inductive bias [10], [24], [25]. Hence, recent DRL studies have focused on introduction to a model of an explicit prior that imposes constraints or regularizations based on the underlying structure of complicated real-world images [26], [27], such as translation and rotation [2], [28], hierarchical features [8], [9], [29] and domain-specific knowledge [10].

Image segmentation has been adopted as an inductive bias in some DRL methods [7], [30], [31]. By assuming that an image consists of local subspaces (sets of pixels) corresponding to objects, DRL methods partition the image to disentangle the latent representation, with the latent variables being separated into distinct sets corresponding to the subspaces. However, conventional DRL methods based on image segmentation change the standard VAE backbone to perform image segmentation and representation learning simultaneously [28]. Since the standard VAE has a statistical backbone supported by the variational Bayesian method, architectural changes of the VAE model may cause unstable learning and deteriorate the ability of disentanglement [5], [28]. In this way, DRL models should have a standard VAE backbone that has preferable DRL properties, whereas it is difficult to perform image segmentation and representation learning in parallel with a VAE model that has the standard structure. The image segmentation task itself has been studied extensively [32]–[34], and it is possible to utilize a pre-trained network to perform image segmentation and concentrate the model only on its latent representation. A pre-trained image segmentation network has remarkable ability to partition an image into object-wise subspaces to disentangle the latent variables while preserving the standard VAE backbone that has favorable DRL properties.

In this paper, we newly introduce an *image segmentation prior* into a VAE model for learning a disentangled representation. An image segmentation prior is an inductive bias that a real-world image consists of subspaces corresponding to objects, which suggests that image segmentation reveals the underlying structure of images. We disentangle the latent representation into each object by imposing this explicit bias with a pre-trained image segmentation network. We develop a drastically simplified DRL method based on the image segmentation prior, which allows a VAE model to focus only on representation learning. We modify the inner structure of the encoder/decoder network to treat segmentation masks explicitly, which enables DRL based on image segmentation while retaining the statistical backbone and the end-to-end learning of the standard VAE. We assume that the entire set of latent variables is composed of two disjoint subsets that represent the foreground and background subspaces, respectively. Then we can implement the partition process with a pre-trained semantic segmentation network, which is a widely studied task in the field of computer vision. In order to retain the original VAE-based end-to-end learning approach, we adopt semantic segmentation instead of instance segmentation or panoptic segmentation. In the encoding/decoding process, the segmentation masks act as alpha mattes [35] for learning a disentangled representation of each subspace. We experimentally validated the effectiveness of our method using real-world image datasets, the CelebA dataset [36] and the Stanford Cars dataset [37].

The contributions of our method can be summarized as follows.

- We propose a novel extended VAE model for DRL in real-world image datasets by introducing an image segmentation network into the encoding process.
- We adopt a pre-trained semantic segmentation network as the prior, which allows our model to focus only on acquiring a disentangled representation without learning image segmentation.

Our model adopts an end-to-end learning approach that retains the preferable ability of the standard VAE and introduces an explicit prior to improve the disentanglement in unsupervised learning.

II. RELATED WORK

A. REPRESENTATION LEARNING

Representation learning is a technology for automatically discovering representations of data that facilitate prediction tasks such as classifications [1]. The performance of machine learning methods is largely dependent on the representation of data [1]. In other words, one of the core tasks in machine learning is how to transform data into simple and discriminative representations. Deep neural networks (DNNs) using multilayer perceptron and convolution operators have achieved remarkable results for classification and estimation tasks in a dataset containing a variety of images [38], and methods using a DNN for representation learning have been widely studied [1], [5].

B. GENERATIVE MODELS

Deep generative models have been extensively applied to representation learning, which aims to learn the simple and low-dimensional representations of complex and high-dimensional data. VAEs [5] and generative adversarial networks (GANs) [39] using DNNs are prominent ones for image data. Although GANs can generate photo-realistic high-quality images, the original GAN [39] does not have an encoder that converts data into latent codes [28]. Hence, we cannot access the latent representation directly, and it is difficult to add constraints explicitly for the latent variables to a GAN-based model [28]. VAE-based deep generative models have attracted much attention in the field of representation learning since a VAE [5] has an encoder and a decoder to learn the mutual transformations of data and latent variables [5], [6], [15].

C. DISENTANGLED REPRESENTATION LEARNING

DRL is one of the most popular topics in studies on representation learning. In a disentangled representation, single latent

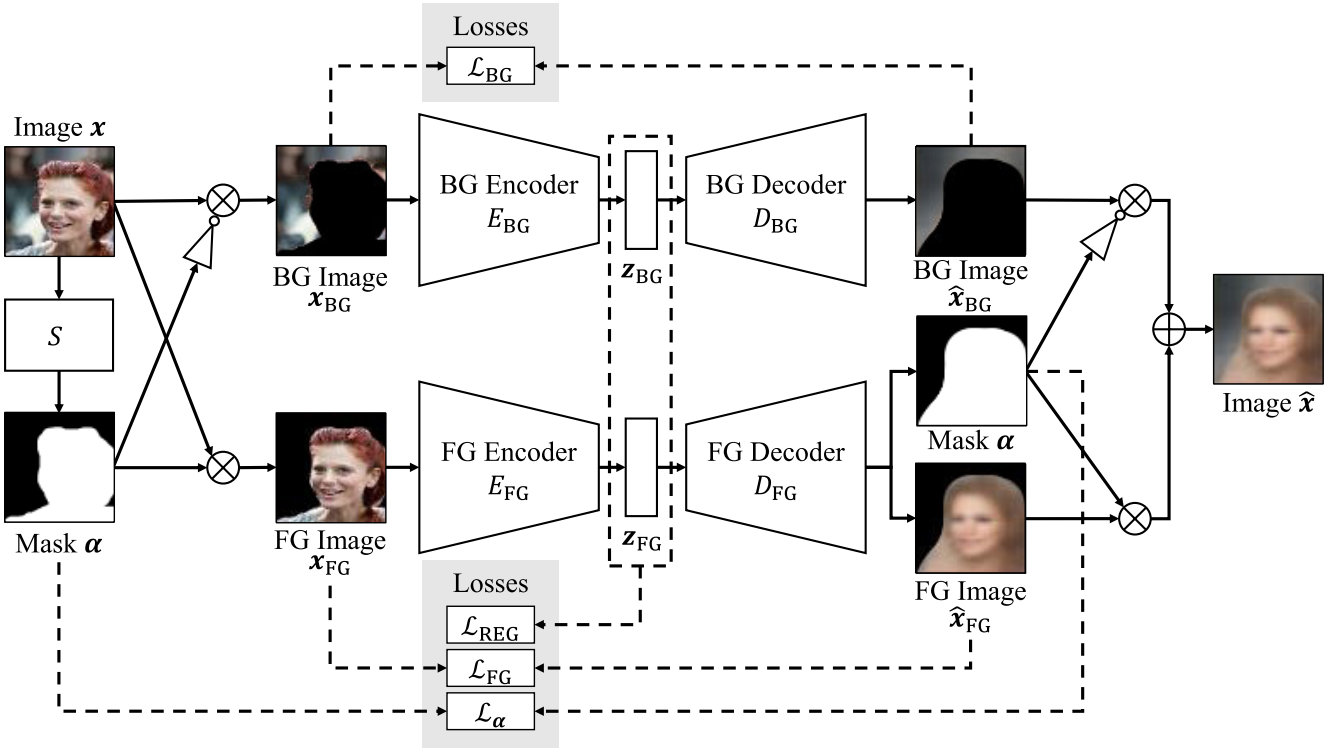


FIGURE 1. Overview of our method. The terms $\mathcal{L}_{\mathrm{BG}}$ and $(\mathcal{L}_{\mathrm{FG}}, \mathcal{L}_{\boldsymbol{\alpha}})$ correspond to the first and second terms in Eq. (10), respectively, and $\mathcal{L}_{\mathrm{REG}}$ means the third and fourth terms in Eq. (10).

variables are sensitive to changes in single factors of variation independently and are relatively invariant to irrelevant factors [1], [6]. There are two main approaches to DRL: implicit disentanglement and explicit disentanglement.

Implicit disentanglement aims to learn disentangled latent representations by regularization so that latent variables are independently distributed. In implicit disentanglement, learning algorithms use only the given data to extract underlying factors of variation and do not use any supervisory signals or explicit inductive biases based on the prior knowledge of a particular domain. β-VAE [6] aims to encourage the latent variables to be independent by introducing a balancing coefficient β to the original VAE. β-VAE with controlled capacity increase (CCβ-VAE) [15] is an extension of β-VAE for solving the problem of information capacity on the basis of the informational bottleneck [15], [40]. β-TCVAE [16] is an extension of β-VAE in which total correlation loss is introduced. However, implicit disentanglement has no guarantee that the obtained latent representations are interpretable [24]. Hence, explicit approaches have been widely studied in the field of disentanglement [26], [27].

Explicit disentanglement is an approach by which the latent variables are explicitly separated into multiple disjoint groups. In the explicit disentanglement, model designers assign roles to the groups of latent variables. The latent representation is disentangled by designing autoencoder networks and a training algorithm so that the latent variables have only information about the roles assigned to their groups. Bepler *et al.* [2] and Detlefsen and Hauberg [3] performed disentanglement on image data by separating the latent variables into two groups assigned to appearance and perspective. Deng *et al.* [10] also developed a disentanglement method for facial images by assigning parameters of a three-dimensional morphable face model [41] to the latent variables.

Disentanglement by segmentation has been a popular topic in the field of explicit DRL for image datasets. This approach is based on the assumption that a real-world image can be split into disjoint regions corresponding to the objects. The segmentation of an image as an inductive bias allows the groups of latent variables to correspond to object regions. Greff *et al.* [30] and Yang *et al.* [31] assigned groups of latent variables corresponding to the objects in an image on the basis of instance segmentation. Awisus *et al.* [7] assigned facial parts (*e.g.,* mouth, nose and hair) to groups of latent variables for face image data.

III. PROPOSED DISENTANGLEMENT METHOD

An overview of our model is shown in Fig. 1. Our model comprises three networks, an encoder E, a decoder D and an image segmentation network S. Our goal is to acquire latent variables $\boldsymbol{z} \in \mathbb{R}^N$ (N being the size of the latent code) that represent an image $\boldsymbol{x} \in [0, 1]^{H \times W \times 3}$ ($H \times W$ being the size of the images). We first obtain an image segmentation mask $\boldsymbol{\alpha} \in [0, 1]^{H \times W \times 1}$ through the image segmentation network S to partition the image $\boldsymbol{x}$ into two subspaces, the foreground subspace $\boldsymbol{x}_{\mathrm{FG}} \in [0, 1]^{H \times W \times 3}$ that contains the main object

and the background subspace $\boldsymbol{x}_{\mathrm{BG}} \in [0,1]^{H\times W\times 3}$ that does not include the main object. Next, we encode the image $\boldsymbol{x}$ into the latent variables $\boldsymbol{z}$ by the encoder E. Finally, we decode the latent variables $\boldsymbol{z}$ into a reconstructed image $\hat{\boldsymbol{x}} \in [0,1]^{H\times W\times 3}$ by the decoder D to duplicate the original image $\boldsymbol{x}$. In Subsection III-A, we introduce the image segmentation prior to the encoder E and the decoder D by utilizing the segmentation mask $\boldsymbol{\alpha}$. In Subsection III-B, we explain the DRL method based on the standard VAE framework.

A. DISENTANGLEMENT VIA IMAGE SEGMENTATION PRIOR

To disentangle the latent space into the two subspaces corresponding to the main object and the background, we partition the image $\boldsymbol{x}$ by the alpha mask $\boldsymbol{\alpha}$. This mask $\boldsymbol{\alpha}$ can be obtained as $\boldsymbol{\alpha} = S(\boldsymbol{x})$ by the pre-trained image segmentation network S. Then we use the mask $\boldsymbol{\alpha}$ to partition the image $\boldsymbol{x}$ into the background image $\boldsymbol{x}_{\mathrm{BG}}$ and the foreground image $\boldsymbol{x}_{\mathrm{FG}}$. This partitioning process can be described as follows:

$$\boldsymbol{x}_{\mathrm{BG}} = \boldsymbol{x} \cdot (1 - \boldsymbol{\alpha}_{\mathrm{tile}}), \tag{1}$$

$$\boldsymbol{x}_{\mathrm{FG}} = \boldsymbol{x} \cdot \boldsymbol{\alpha}_{\mathrm{tile}}, \tag{2}$$

where each operation is element-wise, and $\boldsymbol{\alpha}_{\mathrm{tile}} \in [0,1]^{H\times W\times 3}$ is the extended and copied $\boldsymbol{\alpha}$ along the third dimension. The two sets of latent variables $\boldsymbol{z}_{\mathrm{BG}} \in \mathbb{R}^{N_{\mathrm{BG}}}$ and $\boldsymbol{z}_{\mathrm{FG}} \in \mathbb{R}^{N_{\mathrm{FG}}}$ correspond to $\boldsymbol{x}_{\mathrm{BG}}$ and $\boldsymbol{x}_{\mathrm{FG}}$, respectively ($N_{\mathrm{BG}} + N_{\mathrm{FG}} = N$). These sets $\boldsymbol{z}_{\mathrm{BG}}$ and $\boldsymbol{z}_{\mathrm{FG}}$ are computed in separated VAEs ($E_{\mathrm{BG}}, D_{\mathrm{BG}}$) and ($E_{\mathrm{FG}}, D_{\mathrm{FG}}, D_{\mathrm{FG},\boldsymbol{\alpha}}$). The encoder networks output the parameters of the variational distributions of the latent variables ($\boldsymbol{\mu}_{\mathrm{BG}}, \boldsymbol{\sigma}_{\mathrm{BG}}$) and ($\boldsymbol{\mu}_{\mathrm{FG}}, \boldsymbol{\sigma}_{\mathrm{FG}}$). The variational parameters are represented by the following equations:

$$(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s) = E_s(\boldsymbol{x}_s) \quad (s \in \{\mathrm{BG}, \mathrm{FG}\}), \tag{3}$$

$$\boldsymbol{z}_s \sim \mathcal{N}(\boldsymbol{\mu}_s, \mathrm{diag}(\boldsymbol{\sigma}_s^2)) \quad (s \in \{\mathrm{BG}, \mathrm{FG}\}). \tag{4}$$

To sample the values from these distributions while preserving the differentiability of the networks, we use the "reparameterization trick" [5] expressed as follows:

$$\boldsymbol{z}_s = \boldsymbol{\mu}_s + \boldsymbol{\sigma}_s \cdot \boldsymbol{\epsilon}_s \quad (s \in \{\mathrm{BG}, \mathrm{FG}\}), \tag{5}$$

where each operation is element-wise, and $\boldsymbol{\epsilon}_{\mathrm{BG}}$ and $\boldsymbol{\epsilon}_{\mathrm{FG}}$ are noises sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N_{\mathrm{BG}}})$ and $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N_{\mathrm{FG}}})$, respectively. Finally, the decoder networks output the reconstructed image $\hat{\boldsymbol{x}}$ from the latent variables by the following equations:

$$\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_{\mathrm{BG}} \cdot \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{x}}_{\mathrm{FG}} \cdot (1 - \hat{\boldsymbol{\alpha}}), \tag{6}$$

where the operations of Eq. (6) are element-wise, and

$$\hat{\boldsymbol{x}}_s = D_s(\boldsymbol{z}_s) \quad (s \in \{\mathrm{BG}, \mathrm{FG}\}), \tag{7}$$

$$\hat{\boldsymbol{\alpha}} = D_{\mathrm{FG},\boldsymbol{\alpha}}(\boldsymbol{z}_{\mathrm{FG}}). \tag{8}$$

Thus, we can obtain the disjoint sets of latent variables ($\boldsymbol{z}_{\mathrm{BG}}, \boldsymbol{z}_{\mathrm{FG}}$) that represent the background and foreground separately by the pre-trained network S and the separated VAEs ($E_{\mathrm{BG}}, D_{\mathrm{BG}}$) and ($E_{\mathrm{FG}}, D_{\mathrm{FG}}, D_{\mathrm{FG},\boldsymbol{\alpha}}$).

B. VAE FRAMEWORK FOR DISENTANGLEMENT

We explain the backbone of VAEs, which performs the disentanglement between the latent variables. The fundamental backbone follows the standard definition in a VAE [5]. The encoder E includes variational parameters $\boldsymbol{\phi}_{\mathrm{BG}}$ and $\boldsymbol{\phi}_{\mathrm{FG}}$ that produce variational distributions $q_{\boldsymbol{\phi}_{\mathrm{BG}}}(\boldsymbol{z}_{\mathrm{BG}}|\boldsymbol{x}_{\mathrm{BG}})$ and $q_{\boldsymbol{\phi}_{\mathrm{FG}}}(\boldsymbol{z}_{\mathrm{BG}}|\boldsymbol{x}_{\mathrm{FG}})$, respectively. The decoder D is parameterized by $\boldsymbol{\theta}$ to reconstruct the image $\boldsymbol{x}$ as in Eq. (6). Maximization of the log likelihood $\log p(\boldsymbol{x})$ is performed by maximizing the evidence lower bound (ELBO) [5] as follows:

$$\mathrm{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}; \boldsymbol{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}|\boldsymbol{z})] - \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})), \tag{9}$$

where KL is Kullback-Leibler divergence. We define the loss function $\mathcal{L}$ by modifiying the reconstruction losses for the image segmentation prior and introducing the coefficient β to the ELBO [6], [15], [16] as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \boldsymbol{x}) = {} & \mathbb{E}_{q_{\boldsymbol{\phi}_{\mathrm{BG}}}(\boldsymbol{z}_{\mathrm{BG}}|\boldsymbol{x}_{\mathrm{BG}})}[-\log p_{\boldsymbol{\theta}_{\mathrm{BG}}}(\hat{\boldsymbol{x}}_{\mathrm{BG}}|\boldsymbol{z}_{\mathrm{BG}})] \\ & + \mathbb{E}_{q_{\boldsymbol{\phi}_{\mathrm{FG}}}(\boldsymbol{z}_{\mathrm{FG}}|\boldsymbol{x}_{\mathrm{FG}})}[-\log p_{\boldsymbol{\theta}_{\mathrm{FG}}}(\hat{\boldsymbol{x}}_{\mathrm{FG}}|\boldsymbol{z}_{\mathrm{FG}})] \\ & + \mathrm{KL}(q_{\boldsymbol{\phi}_{\mathrm{BG}}}(\boldsymbol{z}_{\mathrm{BG}}|\boldsymbol{x}_{\mathrm{BG}})||p(\boldsymbol{z}_{\mathrm{BG}})) \\ & + \beta\mathrm{KL}(q_{\boldsymbol{\phi}_{\mathrm{FG}}}(\boldsymbol{z}_{\mathrm{FG}}|\boldsymbol{x}_{\mathrm{FG}})||p(\boldsymbol{z}_{\mathrm{FG}})),\end{aligned} \tag{10}$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\mathrm{BG}}, \boldsymbol{\theta}_{\mathrm{FG}}]$ and $\boldsymbol{\phi} = [\boldsymbol{\phi}_{\mathrm{BG}}, \boldsymbol{\phi}_{\mathrm{FG}}]$. We multiply the coefficient β only to the regularization term for the foreground region containing the main object due to the assumption that real-world images have more essential information in the main object, rather than the background. Assuming that the posterior distributions $p_{\boldsymbol{\theta}_{\mathrm{BG}}}(\hat{\boldsymbol{x}}_{\mathrm{BG}}|\boldsymbol{z}_{\mathrm{BG}})$, $p_{\boldsymbol{\theta}_{\mathrm{FG}}}(\hat{\boldsymbol{x}}_{\mathrm{FG}}|\boldsymbol{z}_{\mathrm{FG}})$ and the prior distributions $p(\boldsymbol{z}_{\mathrm{BG}})$, $p(\boldsymbol{z}_{\mathrm{FG}})$ are Gaussian, the loss function $\mathcal{L}$ is expressed as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \boldsymbol{x}) = {} & \|\boldsymbol{x}_{\mathrm{BG}} - \hat{\boldsymbol{x}}_{\mathrm{BG}}\|^2 + \|\boldsymbol{x}_{\mathrm{FG}} - \hat{\boldsymbol{x}}_{\mathrm{FG}}\|^2 \\ & + \frac{1}{2}\sum_{k=1}^{N_{\mathrm{BG}}}\left[\mu_{\mathrm{BG},k}^2 + \sigma_{\mathrm{BG},k}^2 - \log\sigma_{\mathrm{BG},k}^2 - 1\right] \\ & + \frac{\beta}{2}\sum_{k=1}^{N_{\mathrm{FG}}}\left[\mu_{\mathrm{FG},k}^2 + \sigma_{\mathrm{FG},k}^2 - \log\sigma_{\mathrm{FG},k}^2 - 1\right].\end{aligned} \tag{11}$$

The terms of the KL divergence can be modified as VAE-based models. Using the Monte Carlo method, the training process tries to find optimized parameter values $\boldsymbol{\phi}_{\mathrm{opt}}$ and $\boldsymbol{\theta}_{\mathrm{opt}}$ by the following equation:

$$(\boldsymbol{\phi}_{\mathrm{opt}}, \boldsymbol{\theta}_{\mathrm{opt}}) = \arg\min_{\boldsymbol{\phi},\boldsymbol{\theta}}\left[\sum_{i=1}^{N}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \boldsymbol{x}^{(i)})\right], \tag{12}$$

where $\boldsymbol{x}^{(i)}$ is the i-th sample in the minibatch training.

One of the advantages of our method is that we can introduce other VAE-based DRL methods by modifying the third and fourth terms in Eq. (10) and Eq. (11), since the image segmentation prior preserves the standard VAE backbone as mentioned in Subsection III-A.

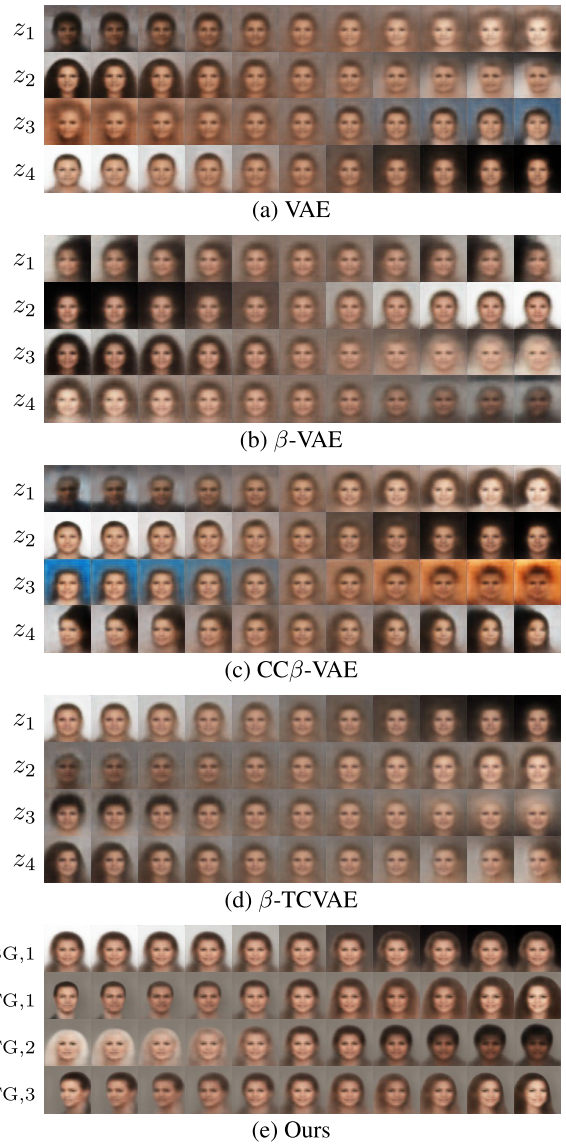


FIGURE 2. Traversal results along latent variables on CelebA. The top four of the N latent variables are selected by their highest variances in the latent codes encoded from the test images.

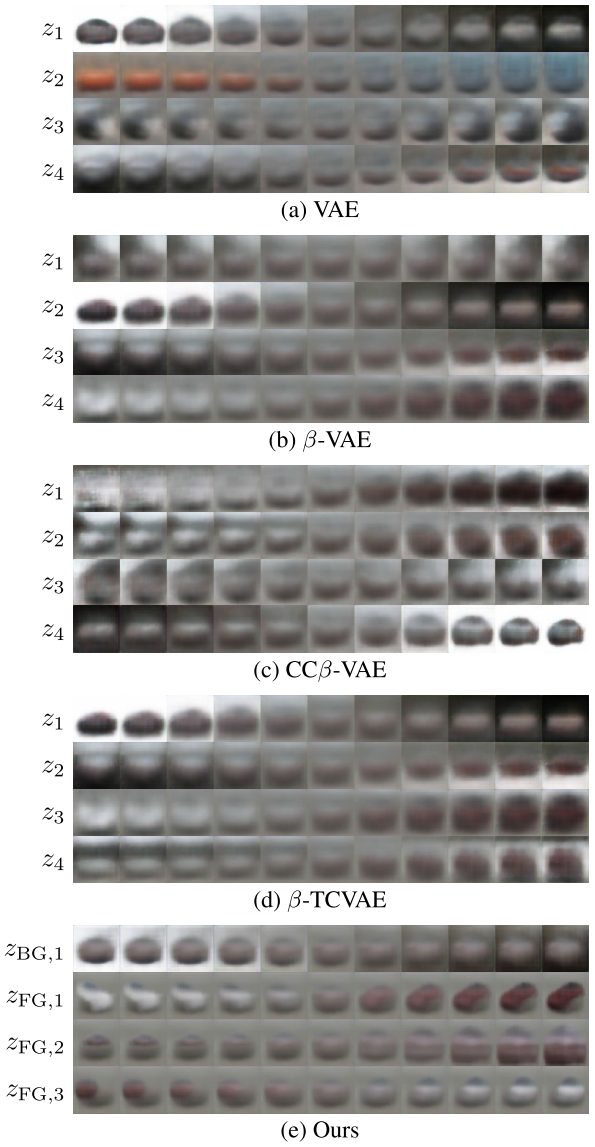


FIGURE 3. Traversal results along latent variables on stanford cars. The top four of the N latent variables are selected by their highest variances in the latent codes encoded from the test images.

IV. EXPERIMENTS

We utilized two real-world image datasets, the CelebA dataset [36] and the Stanford Cars dataset [37]. In the experiments, we used consistent experimental settings in different models as follows.

- We used 202,599 images in the CelebA dataset [36] and used 16,185 images in the Stanford Cars dataset [37]. The images were resized to 64×64 pixels. We split the datasets to 80%/10%/10% for the training/validation/test set, respectively. We performed hyperparameter tuning with 5-fold cross validation, and we computed the quantitative results using the test set.
- We selected the value of the hyperparameter β from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ by the highest WINDIN score in each of the experiments, methods and datasets.
- We set the bottleneck size to $N = 64$ and the number of latent variables $\boldsymbol{z} \in \mathbb{R}^N$. In our model, we assigned the latent variables to the foreground and background as $N_{\mathrm{FG}} = 48$ and $N_{\mathrm{BG}} = 16$ respectively, where $\boldsymbol{z}_{\mathrm{FG}} \in \mathbb{R}^{N_{\mathrm{FG}}}$ and $\boldsymbol{z}_{\mathrm{BG}} \in \mathbb{R}^{N_{\mathrm{BG}}}$. We empirically chose these values of N_{FG} and N_{BG} as $N_{\mathrm{FG}} > N_{\mathrm{BG}}$ because the images were expected to have meaningful contents in the foreground region containing the main object.
- We used the PSPNet-101[1] [34] trained on the Pascal VOC 2012 dataset [42] as the pre-trained image segmentation network S to obtain alpha masks $\boldsymbol{\alpha}$. The Pascal

[1] https://github.com/divamgupta/image-segmentation-keras

VOC 2012 dataset comprises real-world images with objects of various classes, and the literature [34] reported that a single PSPNet achieved remarkable results for the semantic segmentation task in the Pascal VOC 2012 dataset.

- We compared our model with recently proposed VAE-based unsupervised DRL models, the standard VAE [5], β-VAE [6], CCβ-VAE [15] and β-TCVAE [16].

A. QUALITATIVE EVALUATION

We present qualitative results by traversing latent variables on the CelebA dataset in Fig. 2 and on the Stanford Cars dataset in Fig. 3. These results show that the reconstructed images of our model have more distinct contours than those of previous VAE-based models. Furthermore, while standard VAE-based models have latent variables that affect both the main object and the background, our model can disentangle the subspace of the main object from the entire image (shown in Fig. 2(e)). In our model, the subsets of the latent variables $\boldsymbol{z}_{\mathrm{FG}}$ and $\boldsymbol{z}_{\mathrm{BG}}$ correspond separately to the main object, such as a person or a vehicle, and the background that is irrelevant to the main object. Changes in $\boldsymbol{z}_{\mathrm{FG}}$ cause variation within the main object subspace, and changes in $\boldsymbol{z}_{\mathrm{BG}}$ produce variation in the remaining subspace. We show that our model can disentangle information of $\boldsymbol{z}_{\mathrm{FG}}$ and $\boldsymbol{z}_{\mathrm{BG}}$ by introducing alpha compositing to the encoding/decoding process and that our model acquires representations about the foreground object that is invariant to the background. The invariance between $\boldsymbol{z}_{\mathrm{FG}}$ and $\boldsymbol{z}_{\mathrm{BG}}$ suggests that the latent traversal over $\boldsymbol{z}_{\mathrm{FG}}$ do not cause changes in the background region. Since β-VAE increases the constraint on the independence of latent variables, the larger value of β relatively decreases the constraint on the reconstruction (''rate-distortion trade-off'' in the literature [40]). The weaker reconstruction constraint causes more blurry reconstruction images, as larger reconstruction errors are allowed by wider posterior distributions $p_\theta(\boldsymbol{x}|\boldsymbol{z})$.

For images in the CelebA dataset, although our model and the previous VAE-based DRL methods obtained a latent representation corresponding to background brightness, the use of the previous methods resulted in entangled representation for factors other than the background brightness. As shown in Fig. 2, all of the methods successfully disentangled the background brightness factor. The brightness factor corresponds to the latent variable z_1 in VAE [5] (Fig. 2(a)), z_2 in β-VAE (Fig. 2(b)), z_2 in CCβ-VAE (Fig. 2(c)), z_1 in β-TCVAE (Fig. 2(d)) and $z_{\mathrm{BG},1}$ in our method (Fig. 2(e)). However, entangled latent representations for factors other than the background brightness were obtained by the previous methods. Changes in personal appearance coincide with changes in background hue in VAE (z_3 in Fig. 2(a)) and CCβ-VAE (z_3 in Fig. 2(c)) for images in the CelebA dataset. Although β-TCVAE has the two latent variables z_2 and z_3 that are invariant to the background (Fig. 2(d)), the latent variable z_4 appears to have an entangled representation. While our model disentangled the background brightness factor ($z_{\mathrm{BG},1}$ in Fig. 2(e)), we did not confirm any latent variables representing the hue of the background in our model. These results suggest that our model mainly encodes foreground factors into the latent variables. It supports our assumption as in Section III that the foreground region contains the main object and holds the essential information of real-world images.

Also, latent representations corresponding to the main object and the background were successfully obtained separately by our method for images in the Stanford Cars dataset, although only the factor of background luminance was disentangled by the previous methods. As shown in Fig. 3, all of the methods succeeded in disentangling the background brightness in a single latent variable. In our model, the foreground latent variables $z_{\mathrm{FG},1}$, $z_{\mathrm{FG},2}$ and $z_{\mathrm{FG},3}$ correspond to the main object and are invariant to the background as shown in Fig. 3(e). However, a part of the latent traversals caused both changes at the main object and the background in the previous unsupervised DRL methods VAE, β-VAE, CCβ-VAE and β-TCVAE (Fig. 3(a)–(d), respectively).

B. QUANTITATIVE EVALUATION

We assessed the disentanglement capacity of our VAE model with the image segmentation prior for images in the CelebA dataset and the Stanford Cars dataset. For images in a real-world image dataset such as CelebA, it is almost impossible to obtain a true generator that synthesizes photo-realistic images from the latent variables corresponding separately to the actual underlying factors of variation since it is the goal of disentanglement itself [1], [24]. In [43], three important properties of disentanglement are defined: *informativeness*, *separability* and *interpretability*. *Informativeness* means that a latent variable z_i has information of the data $\boldsymbol{x}$ [43]. *Separability* means that two latent variables z_i and z_j ($i \neq j$) do not share common information of the data $\boldsymbol{x}$ [43]. *Interpretability* means that a latent variable z_i only contains information about the pre-defined factor y_k [43]. We utilized the following standard disentanglement metrics that are defined by these three properties and do not require a ground truth generator: WINDIN score, RMIG score and JEMMIG score [43].

As shown in Table 1, our model outperformed other unsupervised VAE-based DRL methods in the WINDIN score. A higher WINDIN value indicates better *informativeness* and *separability* [43]. The WINDIN score is defined by the following equation:

$$\mathrm{WINDIN} = \sum_{i=1}^{N} \rho_i I(\boldsymbol{x}, z_i|\boldsymbol{z}_{\neq i}), \tag{13}$$

where we denote I as mutual information, we denote $\boldsymbol{z}_{\neq i}$ as the set of all latent variables except z_i, and

$$\rho_i = \frac{I(\boldsymbol{x}, z_i)}{\sum_{j=1}^{N} I(\boldsymbol{x}, z_j)}. \tag{14}$$

A high WINDIN value suggests that each of the latent variables z_i has a large amount of information about the image

TABLE 1. Disentanglement. WINDIN score, RMIG score and JEMMIG score over unsupervised DRL methods for images in the CelebA dataset and the stanford cars dataset. (↑ means that higher is better and ↓ means that lower is better. The ranges denote (Mean) ± (Standard error of the mean) from 5-fold cross validation.)

Model	Dataset: CelebA [36]			
	β	WINDIN ↑	RMIG ↑	JEMMIG ↓
VAE [5]	1	0.0150 ± 0.0002	0.01295 ± 0.00057	0.9036 ± 0.0061
β-VAE [6]	10	0.2945 ± 0.0033	0.00720 ± 0.00008	0.9627 ± 0.0086
CCβ-VAE [15]	100	0.0203 ± 0.0002	$\mathbf{0.02577 \pm 0.00029}$	$\mathbf{0.3462 \pm 0.0096}$
β-TCVAE [16]	10	0.1325 ± 0.0024	0.01004 ± 0.00019	0.9582 ± 0.0075
Ours	100	$\mathbf{0.4595 \pm 0.0077}$	0.00284 ± 0.00004	0.9736 ± 0.0027

Model	Dataset: Stanford Cars [37]			
	β	WINDIN ↑	RMIG ↑	JEMMIG ↓
VAE [5]	1	0.0170 ± 0.0001	0.00259 ± 0.00003	1.3434 ± 0.0049
β-VAE [6]	10	0.2290 ± 0.0028	0.00244 ± 0.00005	1.3667 ± 0.0044
CCβ-VAE [15]	100	0.0650 ± 0.0081	0.00261 ± 0.00002	$\mathbf{0.8089 \pm 0.0057}$
β-TCVAE [16]	10	0.0866 ± 0.0037	$\mathbf{0.00294 \pm 0.00007}$	1.3427 ± 0.0067
Ours	100	$\mathbf{0.3765 \pm 0.0077}$	0.00119 ± 0.00002	1.3789 ± 0.0031

TABLE 2. Representation transferability. Estimation error over unsupervised DRL methods for images in the CelebA dataset. (↓ means that lower is better. The ranges denote (Mean) ± (Standard error of the mean) from 5-fold cross validation. Dummy means a dummy classifier that outputs a constant value.)

Model	Estimation Error ↓
Dummy (most frequent value)	19.48%
VAE [5]	$19.09\% \pm 0.047\%$
β-VAE [6] (β=10)	$18.55\% \pm 0.010\%$
CCβ-VAE [15] (β=100)	$17.64\% \pm 0.011\%$
β-TCVAE [16] (β=10)	$18.36\% \pm 0.008\%$
Ours (β=10)	$\mathbf{17.40\% \pm 0.037\%}$

$\boldsymbol{x}$ that the other latent variables $\boldsymbol{z}_{\neq i}$ do not share with the latent variable z_i [43]. In the RMIG score and the JEMMIG score, although CCβ-VAE surpasses other VAE-based generative models, our model sustains the level of the baseline VAE. RMIG and JEMMIG are metrics for *interpretability*: a higher RMIG score means that single latent variables represent the pre-defined factors, and a lower JEMMIG score indicates that each latent variable represents only one of the factors [43]. The RMIG score for the factor y_k is defined as follows:

$$\mathrm{RMIG}_k = I(z_{i^*}, y_k) - I(z_{j^\circ}, y_k), \tag{15}$$

where we denote z_{i^*} and z_{j° as the latent variables with the highest and the second highest mutual information values for y_k, respectively. A high RMIG score suggests that one single latent variable has abundant information about the factor y_k. The JEMMIG score for the factor y_k is defined as follows:

$$\begin{aligned}\mathrm{JEMMIG}_k &= H(z_{i^*}, y_k) - I(z_{i^*}, y_k) + I(z_{j^\circ}, y_k) \\ &= H(z_{i^*}, y_k) - \mathrm{RMIG}_k,\end{aligned} \tag{16}$$

where we denote H as Shannon entropy. A low JEMMIG score means that one single latent variable z_{i^*} represents one single factor y_k [43]. If the latent variable z_{i^*} also represents some of the other factors $\boldsymbol{y}_{\neq k}$, the joint entropy value $H(z_{i^*}, y_k)$ rises due to the additional information that y_k do not share with the latent variable z_{i^*}, which results in the increase of the JEMMIG value. The RMIG score and the JEMMIG score over all of the factors $\boldsymbol{y}$ are reported by the mean values as the following equations:

$$\mathrm{RMIG} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{RMIG}_k, \tag{17}$$

$$\mathrm{JEMMIG} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{JEMMIG}_k, \tag{18}$$

where we denote K as the size of factors $\boldsymbol{y}$. The results show that our model drastically improves *informativeness* and *separability* and preserves *interpretability* of the standard VAE model in the DRL task. The improved *separability* results support the first assumption that a real-world image can be split into object-wise disjoint regions since our method obtains disentangled representations that generate the background and foreground image separately. However, for images in both datasets, our model did not outperform CCβ-VAE and β-TCVAE in terms of the interpretability scores RMIG and JEMMIG. These results suggest a limitation of our method that performs element-wise disentanglement on the basis of the loss function of β-VAE (See Eq. (10)), although our model can explicitly disentangle the groups of latent variables corresponding to the main object and the background as shown in Subsection III-A.

To evaluate representation transferability, we performed estimation of attributes from the latent-space embedding of our model for images in the CelebA dataset. Concretely, we trained a linear classifier $f(\boldsymbol{z}) \simeq \boldsymbol{y}$ that takes the values of the latent variables $\boldsymbol{z}$ as the input and outputs the label

TABLE 3. Estimation on CelebA with different latent variable subsets as the estimation input. Estimation error over our model using different subsets of the latent variables. (↑ means that higher is better and ↓ means that lower is better. The ranges denote (Mean) ± (Standard Error of the Mean) from 5-fold cross validation. None means a dummy classifier that outputs the most frequent value without any inputs.)

Model: Ours (β=100)	
Input	Estimation Error
None	19.48%
$\boldsymbol{z}$	$17.40\% \pm 0.037\%$
$\boldsymbol{z}_{\mathrm{FG}}$ ↓	$17.40\% \pm 0.040\%$
$\boldsymbol{z}_{\mathrm{BG}}$ ↑	$19.48\% \pm 0.031\%$

TABLE 4. Ablation study on alpha compositing. Comparison over our method and that using binary masks instead of alpha compositing in the CelebA dataset. (↑ means that higher is better. The ranges denote (Mean) ± (Standard error of the mean) from 5-fold cross validation.)

Model: Ours (β=100)		
	$\boldsymbol{\alpha} \in$	WINDIN ↑
Binary mask	$\{0,1\}^{H\times W\times 1}$	0.0247 ± 0.0017
Alpha mask (ours)	$[0,1]^{H\times W\times 1}$	$\mathbf{0.4595 \pm 0.0077}$

information $\boldsymbol{y}$. The values of the latent variables $\boldsymbol{z}$ were obtained by encoding the test images with the encoder. We utilized the 40 attributes of the CelebA dataset as the label information $\boldsymbol{y}$. We compared the estimation errors between $\boldsymbol{y}$ and $f(\boldsymbol{z})$ of our model against those of various DRL models. Table 2 shows that our model achieved the best performance in the comparative unsupervised VAE-based DRL methods and reduced the estimation error to 17.40%, which is 1.69 percentage points lower than the baseline VAE. With these results, we can see that the learned representation $\boldsymbol{z}$ by our model is more functional for estimating the label information than existing methods.

Moreover, we examined the different properties of the two disjoint subsets of the latent variables $\boldsymbol{z}_{\mathrm{BG}}$ and $\boldsymbol{z}_{\mathrm{FG}}$. We constructed linear classifiers $f_{\mathrm{BG}}(\boldsymbol{z}_{\mathrm{BG}})$ and $f_{\mathrm{FG}}(\boldsymbol{z}_{\mathrm{FG}})$ corresponding to the subsets $\boldsymbol{z}_{\mathrm{BG}}$ and $\boldsymbol{z}_{\mathrm{FG}}$, respectively. We compared the estimation errors based on the main object information $\boldsymbol{z}_{\mathrm{FG}}$, the remaining information $\boldsymbol{z}_{\mathrm{BG}}$ and the entire information $\boldsymbol{z}$. Table 3 shows that our model acquires a more useful representation for the estimation task in $\boldsymbol{z}_{\mathrm{FG}}$ than in $\boldsymbol{z}_{\mathrm{BG}}$. These results support the assumption that the main object should have sufficient information for estimating the attributes.

To validate our utilization of alpha compositing, we introduce binary masks to our method instead. We compared learned representations from alpha masks and binary masks in Table 4. These results show the efficiency of alpha compositing that retains the differentiability of deep neural networks in our method.

Also, there is still room for discussion about the bottleneck size and allocation ratio between N_{FG} and N_{BG}. Results are expected to be largely dependent on the complexity of datasets. We will study the bottleneck size and allocation into the foreground and background in our future work.

V. CONCLUSION

We have proposed a novel DRL method using an image segmentation prior by a VAE-based generative model for challenging real-world image datasets. Our model retains the standard VAE backbone, which allows us to introduce image segmentation as an inductive bias while preserving the DRL properties of the standard VAE. Qualitative experiments indicated that our model obtains simple, meaningful latent representation that separately denotes the disjoint regions for images in real-world image datasets. Quantitative experiments showed that our model achieves improvements in the disentanglement of latent variables and the transferability of latent representation. However, there is still room for improvement in the element-wise disentanglement of our method. We will consider introducing additional constraints or biases for the disentanglement within each separated group of latent variables.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[2] T. Bepler, E. Zhong, K. Kelley, E. Brignole, and B. Berger, "Explicitly disentangling image content from translation and rotation with spatial-VAE," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 15435–15445.

[3] N. S. Detlefsen and S. Hauberg, "Explicit disentanglement of appearance and perspective in generative models," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 1018–1028.

[4] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 2172–2180.

[5] P. D. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–13.

[6] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–22.

[7] M. Awiszus, H. Ackermann, and B. Rosenhahn, "Learning disentangled representations via independent subspaces," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.

[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

[10] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and controllable face image generation via 3D imitative-contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5154–5163.

[11] O. Press, T. Galanti, S. Benaim, and L. Wolf, "Emerging disentanglement in auto-encoder based unsupervised image content transfer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–25.

[12] C. Eom and B. Ham, "Learning disentangled representation for robust person re-identification," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 5297–5308.

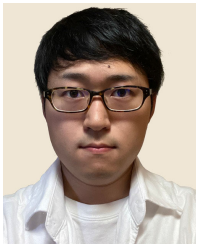
[13] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10257–10266.

[14] M. Willetts, A. Camuto, T. Rainforth, S. Roberts, and C. Holmes, "Improving VAEs' robustness to adversarial attack," 2019, *arXiv:1906.00230*. [Online]. Available: http://arxiv.org/abs/1906.00230

[15] P. Christopher Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," in *Proc. Neural Inf. Process. Syst. (NeurIPS) Workshop*, 2018, pp. 1–11.

[16] T. Q. R. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 2610–2620.

[17] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Balancing learning and inference in variational autoencoders," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5885–5892.
[18] H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, "Learning representations by maximizing mutual information across views," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 15535–15545.
[19] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–16.
[20] S. Zhao, J. Song, and S. Ermon, "Learning hierarchical features from generative models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 4091–4099.
[21] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," in *Proc. Neural Inf. Process. Syst. (NeurIPS) Workshop Bayesian Deep Learn.*, 2018, pp. 1–25.
[22] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. (2017). *dSprites: Disentanglement Testing Sprites Dataset*. [Online]. Available: https://github.com/deepmind/dsprites-dataset/
[23] C. Burgess and H. Kim. (2018). *3D Shapes Dataset*. [Online]. Available: https://github.com/deepmind/3d-shapes/
[24] F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 4114–4124.
[25] B. Dai and D. Wipf, "Diagnosing and enhancing VAE models," 2019, *arXiv:1903.05789*. [Online]. Available: http://arxiv.org/abs/1903.05789
[26] F. Locatello, B. Poole, G. Raetsch, B. Schölkopf, O. Bachem, and M. Tschannen, "Weakly-supervised disentanglement without compromises," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 6348–6359.
[27] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole, "Weakly supervised disentanglement with guarantees," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–36.
[28] Z. Ding, Y. Xu, W. Xu, G. Parmar, Y. Yang, M. Welling, and Z. Tu, "Guided variational autoencoder for disentanglement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7920–7929.
[29] B. Tong, C. Wang, M. Klinkigt, Y. Kobayashi, and Y. Nonaka, "Hierarchical disentanglement of discriminative latent features for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11467–11476.
[30] K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner, "Multi-object representation learning with iterative variational inference," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 2424–2433.
[31] Y. Yang, Y. Chen, and S. Soatto, "Learning to manipulate individual objects in an image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6558–6567.
[32] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 9351, 2015, pp. 234–241.
[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[35] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2970–2979.
[36] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[37] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1097–1105.
[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.
[40] A. A. Alemi, I. Fischer, V. J. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–19.
[41] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 1999, pp. 187–194.
[42] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. Accessed: Oct. 20, 2020. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
[43] K. Do and T. Tran, "Theory and evaluation metrics for learning disentangled representations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–30.



NAO NAKAGAWA (Graduate Student Member, IEEE) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2021, where he is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology. His research interest includes machine learning and its applications.

REN TOGO (Member, IEEE) received the B.S. degree in health sciences from Hokkaido University, Japan, in 2015, and the M.S. and Ph.D. degrees with the Graduate School of Information Science and Technology, Hokkaido University in 2017 and 2019, respectively. He is also a Radiological Technologist. He is currently a Specially Appointed Assistant Professor with the Education and Research Center for Mathematical and Data Science, Hokkaido University. His research interest includes machine learning and its applications.

TAKAHIRO OGAWA (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008. He is currently an Associate Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests include AI, the IoT, and big data analysis for multimedia signal processing and its applications. He is a member of ACM, IEICE, and ITE. He was the Special Session Chair of IEEE ISCE2009, the Doctoral Symposium Chair of ACM ICMR2018, the Organized Session Chair of IEEE GCCE2017-2019, the TPC Vice Chair of IEEE GCCE2018, and the Conference Chair of IEEE GCCE2019. He has been also an Associate Editor of *ITE Transactions on Media Technology and Applications*.

MIKI HASEYAMA (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor, in 1994. She was a Visiting Associate Professor at Washington University, USA, from 1995 to 1996. She is currently a Professor with the Faculty of Information Science and Technology, Division of Media and Network Technologies, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She is a member of the IEICE, Institute of Image Information and Television Engineers (ITE), and Acoustical Society of Japan (ASJ). She has been the Vice-President of ITE, Japan, the Editor-in-Chief of *ITE Transactions on Media Technology and Applications*, and the Director of the International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE).

IS DISENTANGLEMENT ENOUGH? ON LATENT REPRESENTATIONS FOR CONTROLLABLE MUSIC GENERATION

Ashis Pati
Center for Music Technology
Georgia Institute of Technology, USA
ashis.pati@gatech.edu

Alexander Lerch
Center for Music Technology
Georgia Institute of Technology, USA
alexander.lerch@gatech.edu

ABSTRACT

Improving *controllability* or the ability to manipulate one or more attributes of the generated data has become a topic of interest in the context of deep generative models of music. Recent attempts in this direction have relied on learning disentangled representations from data such that the underlying factors of variation are well separated. In this paper, we focus on the relationship between disentanglement and controllability by conducting a systematic study using different supervised disentanglement learning algorithms based on the Variational Auto-Encoder (VAE) architecture. Our experiments show that a high degree of disentanglement can be achieved by using different forms of supervision to train a strong discriminative encoder. However, in the absence of a strong generative decoder, disentanglement does not necessarily imply controllability. The structure of the latent space with respect to the VAE-decoder plays an important role in boosting the ability of a generative model to manipulate different attributes. To this end, we also propose methods and metrics to help evaluate the quality of a latent space with respect to the afforded degree of controllability.

1. INTRODUCTION

Automatic music generation using machine learning has seen significant improvements over the last decade. Deep generative models relying on neural networks have been successfully applied to several different music generation tasks, e.g., monophonic music generation consisting of a single melodic line [1–3], polyphonic music generation involving several different parts or instruments [4,5], and creating musical renditions with expressive timing and dynamics [6, 7]. However, such models are usually found lacking in two critical aspects: *controllability* and *interactivity* [8]. Most of the models typically work as black-boxes, i.e., the intended end-user has little to no control over the generation process. Additionally, they do not allow any modes for interaction, i.e., the user cannot selectively modify the generated music or some of its parts based on desired musical characteristics. Consequently, there have been considerable efforts focusing on controllable music generation [9–11] in interactive settings [12–14]. One promising avenue for enabling controllable music generation stems from the field of representation learning.

Representation learning involves automatic extraction of the underlying factors of variation in given data [15]. The majority of the current state-of-the-art machine learning-based methods aim at learning compact and useful representations [16, 17]. These have been used for solving different types of discriminative or generative tasks spanning several domains such as images, text, speech, audio, and music. A special case of representation learning deals with *disentangled* representations, where individual factors of variation are clearly separated such that changes to a single underlying factor in the data lead to changes in a single factor of the learned disentangled representation [18]. Specifically, in the context of music, disentangled representations have been used for a wide variety of music generation tasks such as rhythm transfer [10, 19], genre transfer [20], instrument rearrangement [21], timbre synthesis [22], and manipulating low-level musical attributes [23–25].

Disentangled representation learning has been an active area of research in the context of deep generative models for music. Previous methods have focused on different types of musical attributes (e.g., note density [23], rhythm [10], timbre [22], genre [20], and arousal [25]) and have achieved promising results. However, contrary to other fields such as computer vision [18, 26], research on disentanglement learning in the context of music has been task-specific and ad-hoc. Consequently, the degree to which disentangled representations can aid controllable music generation remains largely unexplored. While we have shown that unsupervised disentanglement learning methods are not suitable for music-based tasks [27], the use of supervised learning methods has not been systematically evaluated.

In this paper, we conduct a systematic study on controllable generation by using supervised methods to learn disentangled representations. We compare the performance of several supervised methods and conduct a series of experiments to objectively evaluate their performance in terms of disentanglement and controllability for music generation. In the context of this paper, *controllability* is defined as the ability of a generative model to selectively, independently, and predictably manipulate one or more attributes (for instance, rhythm, scale) of the generated data. We show

that while supervised learning methods can achieve a high degree of disentanglement in the learned representation, not all methods are equally useful from the perspective of controllable generation. The degree of controllability depends not only on the learning methods but also on the musical attribute to be controlled. In order to foster reproducibility, the code for the conducted experiments is available online. [1]

2. METHOD & EXPERIMENTAL SETUP

The primary goal of this paper is to investigate the degree to which learning disentangled representations can provide control over manipulating different attributes of the generated music. To this end, we train generative models based on Variational Auto-Encoders (VAEs) [28] to map high-dimensional data in $\mathcal{X}$ to a low-dimensional latent space $\mathcal{Z}$ by approximating the posterior distribution $q(\mathbf{z}|\mathbf{x})$ (encoder). The latent vectors $\mathbf{z} \in \mathcal{Z}$ can then be sampled to generate new data in $\mathcal{X}$ using the learned likelihood $p(\mathbf{x}|\mathbf{z})$ (decoder). We use different supervised learning methods to enforce disentanglement in the latent space by regularizing specific attributes of interest along certain dimensions of the latent space. These attributes can then be manipulated by using simple traversals across the regularized dimensions. Once the models are trained, different experiments are conducted to evaluate disentanglement and controllability.

2.1 Learning Methods

Three different disentanglement learning methods are considered. Each method adds a supervised regularization loss to the VAE-training objective

$$L = L_{\text{VAE}} + \gamma L_{\text{reg}}, \tag{1}$$

where L, L_{VAE}, L_{reg} correspond to the overall loss, the VAE-loss [28], and the regularization loss respectively. The hyperparameter γ is called the regularization strength.

The first method, referred to as I-VAE, is based on the regularization proposed by Adel et al. [29]. It uses a separate linear classifier attached to each regularized dimension to predict the attribute classes. Note that while Adel et al. use this regularization while learning a non-linear transformation of a latent space, we apply it during training of the latent space itself. This is a suitable choice for categorical attributes and is similar to the regularizer used in MIDI-VAE [20]. The second method is the S2-VAE [26]. This regularization, designed for continuous attributes, uses a binary cross-entropy loss to match attribute values to the regularized dimension. The third method is the AR-VAE [30], which uses a batch-dependent regularization loss to encode continuous-valued attributes along specific dimensions of the latent space. This method is effective at regularizing note density and rhythm-based musical attributes [25]. For comparison, baseline results obtained using the unsupervised β-VAE method [31] are also provided.

[1] https://github.com/ashispati/dmelodies_controllability last accessed: 1st Aug 2021

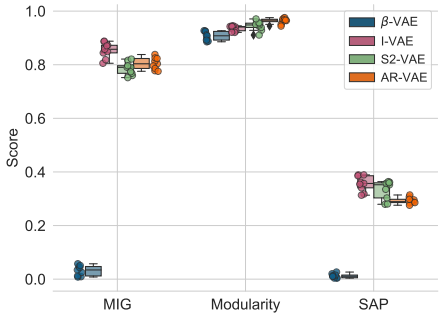


Figure 1: Overall disentanglement performance (higher is better) of different supervised methods on dMelodies. Individual points denote results for different hyperparameter and random seed combinations.

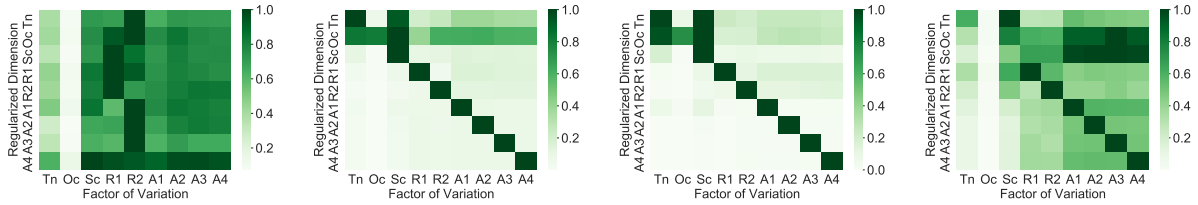
2.2 Dataset & Data Representation

To conduct a systematic study and objectively evaluate the different methods, not only do we need to be able to measure the degree of disentanglement in the learned representations, but we should also be able to measure the attribute values in the generated data. Considering this, we use the dMelodies dataset [27] which is an algorithmically constructed dataset with well-defined factors of variation specifically designed to enable objective evaluation of disentanglement learning methods for musical data. This dataset consists of simple 2-bar monophonic melodies which are based on arpeggiations over the standard I-IV-V-I cadence chord pattern. The dataset has the following factors of variation: *Tonic*, *Octave*, *Scale*, *Rhythm* for bars 1 and 2, and the *Arpeggiation* directions for each of four chords. We use the tokenized data representation used by dMelodies [27].

2.3 Model Architectures & Training Specifications

The VAE architecture is based on a hierarchical RNN model [27], which is inspired by the MusicVAE model [1]. Additional experiments using a CNN-based architecture are omitted here for brevity but provided in the supplementary material. Since both S2-VAE and AR-VAE are designed for continuous attributes, the factors of variation are treated as continuous values by considering the index of the category as the attribute value and then normalizing them to $[0, 1]$. For instance, the *Scale* attribute has 3 distinct options and hence, the normalized continuous values are $[0, \frac{1}{2}, 1]$ corresponding to the major, harmonic minor, and blues scales, respectively. Three different values of regularization strength $\gamma \in \{0.1, 1.0, 10.0\}$ are used.

For each of the above methods and hyperparameter combinations, three models with different random seeds are trained. The dataset is divided into training, validation, and test set using a 70%-20%-10% split. To ensure consistency across training, all models are trained with a batch size of 512 for 100 epochs. The ADAM optimizer [32] is used with a fixed learning rate of $1\text{e}{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\text{e}{-8}$.



(a) β-VAE (b) I-VAE (c) S2-VAE (d) AR-VAE

Figure 2: Attribute-change matrices for different methods. Tn: Tonic, Oc: Octave, Sc: Scale, R1 and R2: rhythm for bars 1 and 2 respectively, A1-A4: arpeggiation direction for the four chords.

3. RESULTS AND DISCUSSION

We now present and discuss the results of the different experiments conducted. The first experiment objectively measures the degree of disentanglement in the representations learned using the different methods. The second experiment evaluates the degree to which each method allows independent control over the different attributes. The third experiment throws additional light into the behavior by visualizing the latent spaces with respect to the different attributes. Then, we introduce a new metric to evaluate the quality of latent spaces with respect to the decoder. Finally, we present a qualitative inspection of the data generated by traversals along different regularized dimensions to further illustrate the key findings.

3.1 Attribute Disentanglement

In order to objectively measure disentanglement, we rely on commonly used metrics: (a) Mutual Information Gap (MIG) [33], which measures the difference of mutual information between a given attribute and the top two dimensions of the latent space that share maximum mutual information with the attribute, (b) Modularity [34], which measures if each dimension of the latent space depends on only one attribute, and (c) Separated Attribute Predictability (SAP) [35], which measures the difference in the prediction error of the two most predictive dimensions of the latent space for a given attribute. For each metric, the mean across all attributes is used for aggregation. For consistency, standard implementations are used [18].

The disentanglement performance of the three supervised methods on the held-out test set is compared against the β-VAE model in Figure 1. Unsurprisingly, all three supervised methods outperform the β-VAE across the three disentanglement metrics. The improvement is much higher for the *MIG* and *SAP* score which both measure the degree to which each attribute is encoded only along a single dimension of the latent space.

Using supervision, therefore, leads to better overall disentanglement. Note that this superior performance is achieved without sacrificing the reconstruction quality. All three supervised methods achieve a reconstruction accuracy $> 90\%$. This is a considerable improvement over the unsupervised learning methods seen in the dMelodies benchmarking experiments (average accuracy of $\approx 50\%$ [30]).

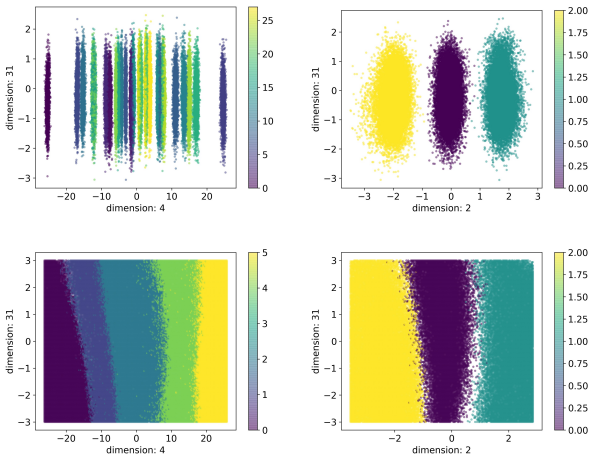
3.2 Independent Control during Generation

Considering that supervised methods can obtain better disentanglement along with good reconstruction accuracy, we now look at how effective these methods are for independently controlling different attributes. To measure this quantitatively, we propose the following protocol. Given a datapoint with latent vector $\mathbf{z}$, 6 different variations are generated by uniformly interpolating along the dimension r_l, where r_l is the regularized dimension for attribute a_l. The limits of interpolation are chosen based on the maximum and minimum latent code values obtained during encoding the validation data. For the β-VAE model, the dimension with the highest mutual information with the attribute is considered as the regularized dimension. An attribute change matrix $A \in \mathbb{R}^{L \times L}$, where L is the number of attributes, is computed using the following formulation:

$$A(m,n) = \sum_{i=1}^{6} \left[0 \neq |a_n(\mathbf{z}_i^m) - a_n(\mathbf{z})| \right], \qquad (2)$$

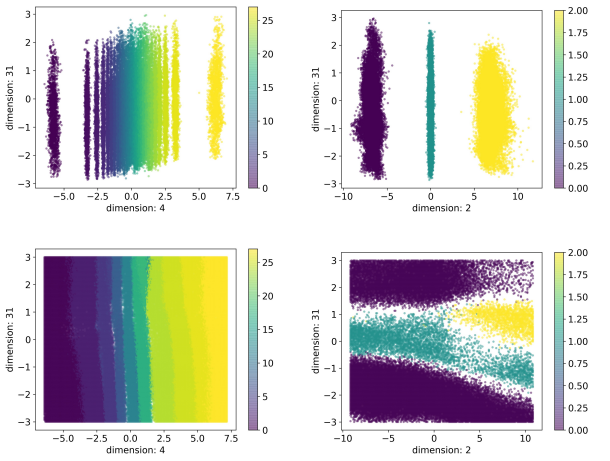
where $A(m,n)$ computes the net change in the n^{th} attribute as one traverses the dimension r_m (which regularizes the m^{th} attribute), $[\cdot]$ represents the inverse Kronecker delta function, $a_n(\cdot)$ is the value of the n^{th} attribute, and $\mathbf{z}_i^m$ is the i^{th} interpolation of $\mathbf{z}$ obtained by traversing along the r_m dimension. This attribute change matrix is computed for each model type by averaging over a total of 1024 data-points in the test-set and across all 3 random seeds (regularization hyperparameters are fixed at $\beta = 0.2$, $\gamma = 1.0$). The matrix is also normalized so that the maximum value across each row corresponds to one. Independent control over attributes should result in the matrix A having high values along the diagonal and low values on the off-diagonal entries which would denote that traversing a regularized dimension only affects the regularized attribute.

The following observations can be made from the matrices visualized in Figure 2. First, β-VAE performs the worst as traversals along different dimensions change multiple attributes simultaneously. Second, among the supervised methods, I-VAE and S2-VAE seem to perform better than AR-VAE. This can be seen from the lighter shades of the off-diagonal elements in the plots for I-VAE and S2-VAE. While the better performance of I-VAE is expected since it is designed for categorical attributes, the poorer performance of AR-VAE in comparison to S2-VAE needs further investigation. Finally, the *scale* attribute (3rd column) changes



(a) *Rhythm Bar 2* (b) *Scale*

Figure 3: Data distribution (top row) and surface plots (bottom row) for I-VAE.



(a) *Rhythm Bar 2* (b) *Scale*

Figure 4: Data distribution (top row) and surface plots (bottom row) for S2-VAE.

the most while traversing the regularized dimensions for the supervised methods. This indicates that all supervised methods struggle in generating notes conforming to particular scales. One explanation for this could be that the *scale* is the most complex among all the attributes. Note that while there is no considerable difference between the disentanglement performance of the three methods (compare Figure 1), I-VAE and S2-VAE show much better performance compared to AR-VAE in this experiment which shows that disentanglement does not ensure better controllability.

3.3 Latent Space Visualization

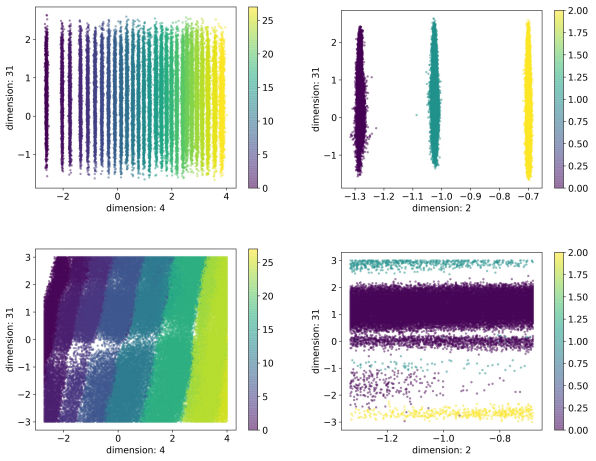
To better understand the difference between disentanglement and controllability of attributes, we try to visualize the structure of the latent space with respect to the different attributes. This is done using 2-dimensional *data distribution* and *latent surface* plots. Both plots show the variance of a given attribute (using different colors for different attribute values) with respect to the regularized dimension (shown on the x-axis) and a randomly chosen non-regularized dimension (shown on the y-axis).

For the data distribution plots, first, latent representations are obtained for data in the held-out test set using the VAE-encoder. Then, for each attribute, these representations are projected onto a 2-dimensional plane where the x-axis corresponds to the regularized dimension and the y-axis corresponds to a non-regularized dimension. To generate the surface plots, for a given attribute, a 2-dimensional plane on the latent space is considered which comprises the regularized dimension for the attribute and a non-regularized dimension. The latent code for the other dimensions is drawn from a normal distribution and kept fixed. The latent vectors thus obtained are passed through the VAE decoder and the attributes of the generated data are plotted.

Figures 3, 4, and 5 show the results for I-VAE, S2-VAE, and AR-VAE respectively. In each figure, the top row corresponds to the data distribution plots, and the bottom row shows the latent surface plots. For the surface plots, the generated data-points sometimes have attribute values that are either not present in the training set or cannot be determined (e.g., the generated melody might not conform to any of the 3 possible scales in the dataset, or the arpeggiation direction might be neither up nor down). These *undefined* or out-of-distribution attribute values are shown as empty spaces in the latent surface plots.

For all three methods, the data distribution plots (top rows) show a clear separation of attribute values along the regularized dimension which explains the high disentanglement performance seen in Section 3.1. However, the methods differ considerably when the latent surface plots (bottom rows) are compared. I-VAE (see Figure 3) shows good performance where moving along the regularized dimension (x-axis) changes the corresponding attribute, while traversals along the non-regularized dimension (y-axis) have little effect. However, the manner of change is unpredictable. For instance, in Figure 3(a)(bottom), only 5 out of the 28 possible rhythms are generated. In addition, the order of the generated rhythms is different from the encoder distribution in Figure 3(a)(top). In contrast, for S2-VAE, the gradual change of color in Figure 4(a)(bottom) shows a high degree of controllability for the rhythm attribute. However, it struggles to control the *scale* attribute. Traversing along the non-regularized dimension in Figure 4(b)(bottom) results in an undesirable change in the *scale* of the generated melody. The latent space of AR-VAE (see Figure 5) has the most discrepancies. Not only is the latent space not centered around the origin (see the top row of Figure 5(b)) for the *scale* attribute, but the degree of controllability is also poor. For instance, the *scale* attribute does not change at all along the regularized dimension (see Figure 5(b)(bottom)). In addition, the empty spaces in the surface plots show that many of the generated data-points have an out-of-distribution attribute value. Results for all other attributes are provided in the supplementary material.

The empty regions in the latent spaces show that while these methods can train strong discriminative encoders



(a) *Rhythm Bar 2* (b) *Scale*

Figure 5: Data distribution (top row) and surface plots (bottom row) for AR-VAE.

which are good for disentanglement, they tend to have weak generative decoders which are incapable of utilizing the learned disentangled representations thereby resulting in *holes* or vacant regions in the latent space where the behavior of the decoder is unpredictable.

3.4 Latent Density Ratio

From the perspective of the VAE-decoder, holes in the latent space can have a significant impact on controllability. Yet, established metrics do not capture this phenomenon properly. To help quantify this, we propose the Latent Density Ratio (LDR) metric. We first sample a set of N (=10k) points in the latent space, pass them through the VAE decoder, and compute the percentage of data-points with valid attribute values out of the total number N. The overall LDR is obtained by averaging this metric across all attributes. The results in Table 1 show that both S2-VAE and I-VAE have a lower degree of holes (higher LDR value) in comparison to AR-VAE which is in line with observations in the previous experiments.

3.5 Qualitative Inspection of Latent Interpolations

Finally, we take a qualitative look at the data generated by the different methods while traversing the latent space along the regularized dimensions. Ideally, traversals along a regularized dimension should only cause changes in the corresponding attribute while leaving the other attributes unchanged. In addition, the regularized attribute should also change in a predictable manner. Figure 6 shows the results for the I-VAE method. For each sub-figure, different rows correspond to melodies generated by traversing along the regularized dimension for the attribute in the sub-figure caption. Results for S2-VAE and AR-VAE are shown in Figures 8 and 7, respectively.

Across methods, most of the time, the melodies generated by traversing along regularized dimensions show changes in the corresponding attribute only. For instance,

Learning Method	LDR
I-VAE	0.448
S2-VAE	0.544
AR-VAE	0.244

Table 1: LDR metric (higher is better) for different methods

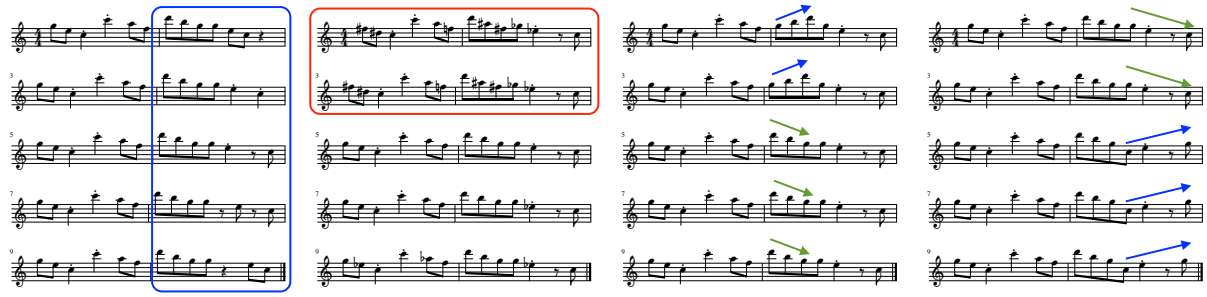
in Figures 6(a) and 7(a), only the rhythm of the second bar changes while the rest of the melody stays intact. In Figure 6(c,d), the arpeggiation directions of the third and fourth chords are flipped, respectively. Also, in Figure 6(b), all the other attributes remain constant (rhythm, arpeggiation directions) while the pitches of the generated notes change to reflect different scales. While this is desirable, there are a few important things to note.

First, the *scale* attribute seems hard to control. For instance, in Figure 6(b), for I-VAE, some of the generated melodies (the first two rows) do not conform to any of the scales present in the dataset. In Figure 7(b), for AR-VAE, the *scale* does not change at all. This difficulty in controlling the *scale* attribute was also observed in Section 3.2. Second, depending on the holes in the latent space, traversals along regularized dimensions sometimes create melodies with attributes that are unseen in the training data. This happens also for attributes other than *scale*. For instance, in Figure 7(c), row 2, the third chord has an unseen arpeggiation direction. Finally, for I-VAE, the direction of change for arpeggiation factors (see Figure 6(c,d)) is unpredictable. While the arpeggiation direction (of the third chord) goes from up to down in Figure 6(c), the direction (for the fourth chord) is flipped from down to up in Figure 6(d). This is due to the I-VAE regularization formulation which is agnostic to the order of the categorical attributes. Contrast this to AR-VAE and S2-VAE, where the nature of the change in the attribute values is predictable. The direction of arpeggiation will always go from up to down for these methods (see Figures 8(a,b) and 7(c,d)).

3.6 Discussion

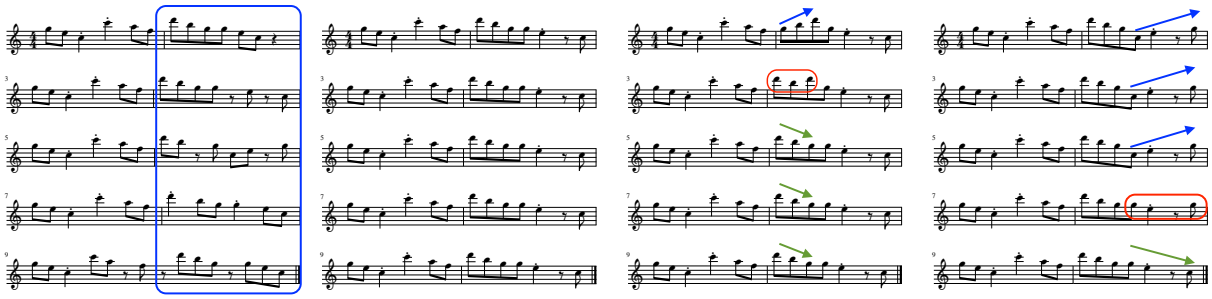
The results of the experiments in this section show that supervised methods for disentanglement perform significantly better than unsupervised methods. This is expected since the former use attribute-specific information during training to guide the model towards learning better representations. Among the supervised methods, there are no major differences in terms of the disentanglement metrics in Section 3.1. However, controllability during data generation (discussed in Sections 3.2, and 3.5) differs considerably between the methods. These differences suggest that while disentanglement is closely related to a strong encoder (learning the posterior $q(\mathbf{z}|\mathbf{x})$), improving controllability requires a strong decoder (learning the likelihood $p(\mathbf{x}|\mathbf{z})$). This explains the often better performance of conditioning-based methods relying on adversarial training of decoders [36, 37].

Visualizing the latent spaces (in Section 3.3) with respect to the attribute values highlights that the presence or



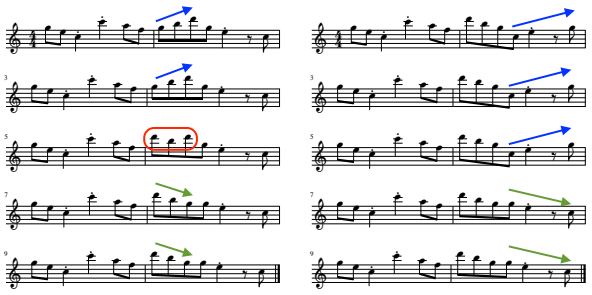
(a) Rhythm Bar 2 (b) Scale (c) Arp Chord 3 (d) Arp Chord 4

Figure 6: Generated data by traversing along regularized dimensions for I-VAE.



(a) Rhythm Bar 2 (b) Scale (c) Arp Chord 3 (d) Arp Chord 4

Figure 7: Generated data by traversing along regularized dimensions for AR-VAE.



(a) Arp Chord 3 (b) Arp Chord 4

Figure 8: Generated data by traversing along regularized dimensions for S2-VAE.

absence of holes in the learned latent space plays a crucial role in the degree of controllability afforded by a model. The LDR metric proposed in Section 3.4 is an attempt to quantify this behavior. Note that other factors can be considered while evaluating controllability that have been left out of this study. For instance, for continuous-valued attributes, one would prefer the regularized dimension having a positive correlation with the attribute value [30].

4. CONCLUSION

In this paper, we present a systematic investigation of the relationship between attribute disentanglement and controllability in the context of symbolic music. Through a diverse set of experiments using different methods, we show that even though different supervised learning techniques can force effective disentanglement in the learned representations to a comparable extent, not all methods are equally effective at allowing control over the attributes during the data generation process. This distinction is important because controllability is paramount for generative models [8] and is often not taken into account while evaluating disentanglement learning methods.

An important observation is the issue of holes in latent spaces. It should be noted this has also been seen in other data domains relying on discrete data such as text [38]. There are a few promising directions to address this problem. One option is to constrain the latent space to conform to a specific manifold and perform manipulations within this manifold [38, 39]. An alternative direction could be to learn specific transformation paths within the existing latent manifold to avoid these holes [40].

The experiments in this paper have used labels from the entire training set. Another interesting direction for future studies could be to extend these experiments to a semi-supervised paradigm by using a limited number of labels obtained from only a fraction of the training set [26]. This would increase the confidence in applying these methods to real-world data where obtaining label information for the entire dataset might be either too costly or simply impossible.

5. ACKNOWLEDGMENTS

The authors would like to thank NVIDIA Corporation (Santa Clara, CA, United States) for supporting this research via the NVIDIA GPU Grant program.

6. REFERENCES

[1] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," in *Proc. of 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweeden, 2018.

[2] F. Colombo, S. Muscinelli, A. Seeholzer, J. Brea, and W. Gerstner, "Algorithmic composition of melodies with deep recurrent neural networks," in *Proc. of 1st Conference on Computer Simulation of Musical Creativity (CSMC)*, Huddersfield, UK, 2016.

[3] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, "Music transcription modelling and composition using deep learning," in *Proc. of 1st Conference on Computer Simulation of Musical Creativity (CSMC)*, Huddersfield, UK, 2016.

[4] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation," in *Proc. of 18th International Society of Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 324–331.

[5] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *Proc. of 29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, 2012.

[6] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," in *Proc. of International Conference of Learning Representations (ICLR)*, New Orleans, USA, 2019.

[7] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: Learning expressive musical performance," *Neural Computing and Applications*, pp. 1–13, 2018.

[8] J.-P. Briot and F. Pachet, "Deep learning for music generation: Challenges and directions," *Neural Computing and Applications*, 2018.

[9] A. Pati, A. Lerch, and G. Hadjeres, "Learning to Traverse Latent Spaces for Musical Score Inpainting," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[10] R. Yang, D. Wang, Z. Wang, T. Chen, J. Jiang, and G. Xia, "Deep music analogy via latent representation disentanglement," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[11] C.-Z. A. Huang, T. Cooijmans, A. Roberts, A. Courville, and D. Eck, "Counterpoint by convolution," in *Proc. of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017.

[12] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: A steerable model for Bach chorales generation," in *Proc. of 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1362–1371.

[13] C. Donahue, I. Simon, and S. Dieleman, "Piano genie," in *Proc. of 24th International Conference on Intelligent User Interfaces (IUI)*, Los Angeles, USA, 2019, pp. 160–164.

[14] T. Bazin and G. Hadjeres, "Nonoto: A model-agnostic web interface for interactive music composition by inpainting," in *Proc. of 10th International Conference on Computational Creativity (ICCC)*, UNC Charlotte, NC, USA, 2019.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 2013.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of 37th International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 1597–1607.

[17] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2020.

[18] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations," in *Proc. of 36th International Conference on Machine Learning (ICML)*, Long Beach, California, USA, 2019.

[19] J. Jiang, G. G. Xia, D. B. Carlton, C. N. Anderson, and R. H. Miyakawa, "Transformer VAE: A Hierarchical Model for Structure-Aware and Interpretable Music Representation Learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 516–520.

[20] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, "MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer," in *Proc.*

of 19th International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 2018.

[21] Y.-N. Hung, I.-T. Chiang, Y.-A. Chen, and Y.-H. Yang, "Musical composition style transfer via disentangled timbre representations," in *Proc. of 28th International Joint Conference on Artificial Intelligence (IJCAI)*, Macao, China, 2020.

[22] Y.-J. Luo, K. Agres, and D. Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[23] G. Hadjeres, F. Nielsen, and F. Pachet, "GLSR-VAE: Geodesic latent space regularization for variational autoencoder architectures," in *Proc. of IEEE Symposium Series on Computational Intelligence (SSCI)*, Hawaii, USA, 2017, pp. 1–7.

[24] A. Pati and A. Lerch, "Latent space regularization for explicit control of musical attributes," in *Proc. of ICML Workshop on Machine Learning for Music Discovery Workshop (ML4MD), Extended Abstract*, Long Beach, California, USA, 2019.

[25] H. H. Tan and D. Herremans, "Music FaderNets: Controllable music generation based on high-level features via low-level feature modelling," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020.

[26] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem, "Disentangling factors of variations using few labels," in *Proc. of 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

[27] A. Pati, S. Gururani, and A. Lerch, "dMelodies: A Music Dataset for Disentanglement Learning," in *Proc. of 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020.

[28] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. of 2nd International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014.

[29] T. Adel, Z. Ghahramani, and A. Weller, "Discovering Interpretable Representations for Both Deep Generative and Discriminative Models," in *Proc. of 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 50–59.

[30] A. Pati and A. Lerch, "Attribute-based Regularization of Latent Spaces for Variational Auto-Encoders," *Neural Computing and Applications*, 2020. [Online]. Available: https://doi.org/10.1007/s00521-020-05270-2

[31] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in *Proc. of 5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[32] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of 3rd International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015.

[33] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating Sources of Disentanglement in Variational Autoencoders," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, Montréal, Canada, 2018.

[34] K. Ridgeway and M. C. Mozer, "Learning Deep Disentangled Embeddings With the F-Statistic Loss," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, Montréal, Canada, 2018, pp. 185–194.

[35] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational Inference of Disentangled Latent Concepts from Unlabeled Observations," in *Proc. of 5th International Conference of Learning Representations (ICLR)*, Toulon, France, 2017.

[36] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader Networks:Manipulating Images by Sliding Attributes," in *Advances in Neural Information Processing Systems 31 (NeurIPS)*, Long Beach, California, USA, 2017, pp. 5967–5976.

[37] L. Kawai, P. Esling, and T. Harada, "Attributes-aware deep music transformation," in *Proc. of 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020.

[38] P. Xu, J. C. K. Cheung, and Y. Cao, "On variational learning of controllable representations for text without supervision," in *Proc. of 37th International Conference on Machine Learning (ICML)*, 2020.

[39] M. Connor and C. Rozell, "Representing Closed Transformation Paths in Encoded Network Latent Space," in *Proc. of 34th AAAI Conference on Artificial Intelligence*, New York, USA, 2020.

[40] D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow, "Understanding and improving interpolation in autoencoders via an adversarial regularizer," in *Proc. of 7th International Conference on Learning Representations (ICLR)*, New Orleans, USA, 2019.

sensors

MDPI

Article

Frequency Disentanglement Distillation Image Deblurring Network

Yiming Liu [1], Jianping Guo [2,*], Sen Yang [1], Ting Liu [1], Hualing Zhou [1], Mengzi Liang [1], Xi Li [1] and Dahong Xu [1]

[1] College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China; liuyiming@hunnu.edu.cn (Y.L.); 202070291671@hunnu.edu.cn (S.Y.); tinaliuting2021@163.com (T.L.); 202020291657@hunnu.edu.cn (H.Z.); 202020291651@hunnu.edu.cn (M.L.); lixi@hunnu.edu.cn (X.L.); xudahong@hunnu.edu.cn (D.X.)

[2] College of Physical Education, Hunan Normal University, Changsha 410081, China

* Correspondence: gjp2009@hunnu.edu.cn; Tel.: +86-139-7515-0758

Abstract: Due to the blur information and content information entanglement in the blind deblurring task, it is very challenging to directly recover the sharp latent image from the blurred image. Considering that in the high-dimensional feature map, blur information mainly exists in the low-frequency region, and content information exists in the high-frequency region. In this paper, we propose a encoder–decoder model to realize disentanglement from the perspective of frequency, and we named it as frequency disentanglement distillation image deblurring network (FDDN). First, we modified the traditional distillation block by embedding the frequency split block (FSB) in the distillation block to separate the low-frequency and high-frequency region. Second, the modified distillation block, we named frequency distillation block (FDB), can recursively distill the low-frequency feature to disentangle the blurry information from the content information, so as to improve the restored image quality. Furthermore, to reduce the complexity of the network and ensure the high-dimension of the feature map, the frequency distillation block (FDB) is placed on the end of encoder to edit the feature map on the latent space. Quantitative and qualitative experimental evaluations indicate that the FDDN can remove the blur effect and improve the image quality of actual and simulated images.

Keywords: image deblurring; feature disentanglement; distillation block; frequency split

Citation: Liu, Y.; Guo, J.; Yang, S.; Liu, T.; Zhou, H.; Liang, M.; Li, X.; Xu, D. Frequency Disentanglement Distillation Image Deblurring Network. *Sensors* **2021**, *21*, 4702. https://doi.org/10.3390/s21144702

Received: 14 June 2021
Accepted: 5 July 2021
Published: 9 July 2021

1. Introduction

Image blur, caused by camera shake, object motion or out-of-focus, is one of the most common visual artifacts. The purpose of image deblurring is to recover a sharp latent image from a blurred image using edge structure and details when a single blurred image is given. Image deblurring has long been an essential task in computer vision and image processing. The image blurring can be expressed by Equation (1). Blurred image, blur kernel, sharp image, noise and the operation of convolution are represented by B, K, I, N and $*$ respectively.

$$B = K * I + N, \tag{1}$$

According to whether the blur kernel is a known condition, the deblurring task can be divided into two major branches: non-blind image deblurring [1] and blind image deblurring [2–10]. For non-blind deblurring, the information of the blur kernel K must be known in advance. According to the blur kernel, the sharp image can be recovered by reverse convolution on the blurred image. This process is relatively easy to follow by a calculation standard. However, in most cases, the blur kernel is unpredictable. The blind image deblurring is to estimate a sharp image when only the blurred image is given. Obviously, It can be seen that non-blind deblurring is an ill-posed problem. The traditional methods [11–13] tend to estimate the blur kernel and the sharp image simutaneously. These

methods generally use assumed prior knowledge to limit the uncertainty of the blur kernels, thus turning the blind deblurring problem into a non-blind problem. For example, Ref. [11] believes blur will cause the gradient change of the image. The sharp image is restored by updating the gradient. Pan et al. [13] proposed a dark channel prior constraining method, which can solve this ill-posed problem. However, due to the complexity of blur kernel in the real world, the assumed prior knowledge is inevitably limited and difficult to fully express the blurry situation in the real scene, which leads to the artifacts in the restored image. Moreover, these methods based on iterative optimization techniques are very computationally expensive and usually require tuning a large number of parameters.

With the development of deep learning, many methods [2–10] began to no longer predict blur kernels but directly restored sharp images end-to-end by constructing an encoder–decoder structure. Encoder–decoder architecture aims to invert a given blur image back into the latent space by the encoder, and the image then can be faithfully reconstructed from the latent feature by the decoder. Nah [2] has built three parallel encoder–decoder paths, and different paths receive different image scales, so that a sharp image can be gradually recovered from the blurred image. SRNnet [3] also uses a similar multi-scale framework, but it introduced the LSTM [14] to share the intermediate layer information in the latent space. Deblurgan [5] uses the unet-based network as the backbone of the generator. Deblurganv2 [9] replaced unet-based with pyramid network to achieve the effect of feature reuse. Through our observation, although the above method solves the problem of image blurring to some extent. They still cannot restore the original information of the image well and even introduce artifacts inevitably. Through our analysis, such defects are mainly due to the design of the encoder–decoder architecture can not disentangle the blur information from the content information. The encoder task is to extract semantic content feature map from the image, which will served as important clues to reconstruct a high-fidelity sharp image. Then, the decoder is guided by the loss function to supplement the detailed information lost due to the down-sampling in the decoder. It seems reasonable that since the original blur information has also been eliminated mainly in the encoding process. There may be only the content information of the image left in the intermediate feature maps of the latent space. In fact, the blur information of the image is entangled with the content information. Even if the encoder extracts the semantic feature of content information to the maximum extent, there will still be some surviving blur information that is entangled with it so that it is impossible to remove blur features as a single vector or a independent feature channel from the whole feature maps through linear reorganization. These entangled blur features are mistaken as valuable clues in the process of decoder, which disturbs the image reconstruction process of the model and leads to produce unnatural textures and artifacts.

In response to the above problems, we proposed a frequency disentanglement distillation image deblurring network (FDDN) edit on intermediate feature map in the latent space. We proposed the frequency split block(FSB), inspired by octconv [15], disentangles the intermediate feature map in latent space in the dimension of frequency. We think that blur information usually exists in low-frequency features, and semantic feature of content information usually exists in high-frequency features. Extracting blur information from blur image or disentangling blur information from content information is a complicated thing. We cannot solve it directly either, but we can narrow down the scope of solution through our frequency distillation block (FDB). FDB will greatly retain the high-frequency features in the feature channel and distillate the low-frequency features so as to solve the entanglement problem of blur information as far as possible.

There are three clear positive impacts of the FDDN algorithm. First, FDDN is the first time to define the deblurring task as the disentanglement operation of deblurring information and image content information, which is different from the previous encoder–decoder algorithm to generate sharp image directly. Compared with the direct method, FDDN is more instructive to the network and purposefully guides the network to eliminate blurry information, rather than relying solely on the training set and loss function. Second, FDDN algorithm has a great positive impacts in network parameters and running speed,

which is due to the FDB disentanglement work at latent space, where the complexity of eigenvectors is the lowest. Third, FDDN has achieved a excellent result both in quantitative and qualitative aspects.

Our contribution can be summarized in the following three points:

- A frequency split block (FSB) is proposed, distilling high-frequency and low-frequency features in different channels.
- We propose a frequency distillation block (FDB) that can better retain the information of the high-frequency characteristic channel and filter and reorganize the information of the low-frequency characteristic channel.
- A lot of experiments have been conducted to prove the validity of the FDDN that we designed.

2. Related Work

Image deblurring has also been rapidly developed because of the multi-scale mechanism [2,3,9,16,17]. Nah et al. [2] creatively proposed the multi-scale deblurring pipline at the first time, which introduced three kinds of blurry images with different sizes into the model, and achieve a state of the art result in the year of 2017. This multi-scale design make the model can perceive both detail and semantic information. However, this method required the network to carry out feature extraction and image reconstruction for three times, which lead to the number of network parameters is too large, and the model performance is low. SRN [3] optimized the pipeline of multi scale by employing LSTM mechanism. This design let model share the feature extraction results across scales. However, the problem of parameter overload cannot be solved fundamentally. Zhang et al. [16] investigate a new scheme which exploits the deblurring cues at different scales via a hierarchical multi-patch model, and propose a simple yet effective multi-level CNNs model called Deep Multi-Patch Hierarchical Network (DMPHN), which uses multi-patch hierarchy as input. Gao et al. [17] believed that [2,3] were in two extremes, in which the information of Deep deblurring net [2] was utterly independent at each scale, while SRN net [3] fused all intermediate information without screening and both of them could only obtain suboptimal results. Therefore, [17] proposed a selective sharing mechanism on the basis of multi-scales and solved the complex problem of very deep network training through jump connection. Deblurganv2 [9] also introduces a FPN network [18] into the generator to take advantage of multi-scale feature information, enabling the integration of high-level semantic information with low-level detail information. However, this design still has some shortcomings. The upper semantic information will be diluted in the process of transmission, so the higher semantic information will be gradually weakened in the process of network transmission. Kupyn et al. [5] creatively introduced the model of adversarial generation network in image deblurring, which define the task of deblurring as a transformer task. The generator received blur images as input and sharp images as output, and the discriminator was used to discriminate the authenticity of the generated images.

In 2021, image deblurring [19–24] also achieved good results. GCResNet [19] proposed a new codec network, in order to increase the amount of convolution of the graph, the feature map is converted into the vertices of the pre-generated graph to synthesize the structure data of the graph. By doing this, we apply Tulaplacian regularization to feature maps to make them more structured. Pan et al. cited 34 to divide the deblurring process into two steps and proposed a two-stage network. In the first stage, a public convolutional network is used to generate an initial deblurred image. In the second stage, the initial data distribution is transformed into a potential sharp image distribution, and sharp edges are obtained through a priori network. In addition, they proposed a relativistic training strategy aimed at learning the priors of potentially sharp images to train prior networks. Wu et al. [21] designed a deblurring method based on a two-stage wavelet-based convolutional neural network, which embeds discrete wavelet transform to separate image context and texture information, and reduces computational complexity. In addition, they modified the initial module by increasing the pixel attention mechanism and the channel

scale factor so that the weight of each convolution kernel was changed, and at the same time, the receiving field was increased and the parameters of the module were significantly reduced. In order to guide the network to perform higher-quality deblurring and improve the feature similarity between the restored image and the clear image, SharpGAN [22] proposes a method that combines feature loss of different levels of image features. In addition, they introduced the network into the receiving domain block network to improve the ability to extract fuzzy image features. Wang et al. [23] proposed a new framework that uses depth variational Bayes to blindly deblur the image. This framework uses discrete reasoning and deep neural networks (DNNs) to jointly estimate the posterior of potential clean images and blur kernels. In addition, under the guidance of the lower bound of evidence, the data of clean images and blur kernels can be considered. Drive a priori supervision and physical fuzzy models to train the inference DNNs involved. MPRNet [24] proposed a method, which has two characteristics. One is that information is exchanged in the order from early to late; the other is to avoid information loss. It is also in the feature processing block. A horizontal connection is added, and a tightly connected multi-level architecture is created on this basis.

3. Proposed Method

3.1. Overview

In this article, we also constructed an encoder–decoder network structure as shown in Figure 1, and it can be divided into three parts. In the first part, the structure of the encoder is mainly composed of two identical inception down-sampling blocks, as shown in Figure 2. The inception down-sampling is inspired by the inceptions block [25], where the down-sampling operation has been completed by max-pooling with a kernel of 2×2 and a stride of 1, so that the length and width of the feature map are each $1/2$ of the original length and width. Before pooling, the feature is resampled with the convolution kernel 3×3 and the stride is 1 without changing the size of the feature map. The receptive field of the feature pixel will expand due to the resampling, thus solving the problem of detail loss in the process of feature map scale reduction. After the encoding path, it enters the second part. The middle layer feature in the latent space completes the operation of frequency disentangle and distillation here through 16 frequency distillation blocks(FDB) to reduce the disturbance of the network by useless features. We will introduce it in detail in Section 3.2. The third part is the decoder, in which the gray module is the residual channel attention block [26], which is also used for feature reorganization, also does not change the size of the input feature map, but through its channel attention mechanism to make the decoding network more targeted recovery Image details. The green module is pixel-shuffle convolution [27], which uses convolution to expand the number of feature channels without changing the size of the feature map. Then squeeze the feature maps of multiple channels into one feature map to achieve the purpose of up-sampling, as shown in Figure 3. Since the expanded feature channels are obtained by convolution layers, these convolutions can be trained together with other parameters, so compared with the traditional interpolation up-sampling method [16,17], it can produce more realistic results. Furthermore, in our frequency disentangle distillation image deblurring network (FDDN), we design a large number of skip connection mechanisms [17] between encoding path and decoding path.

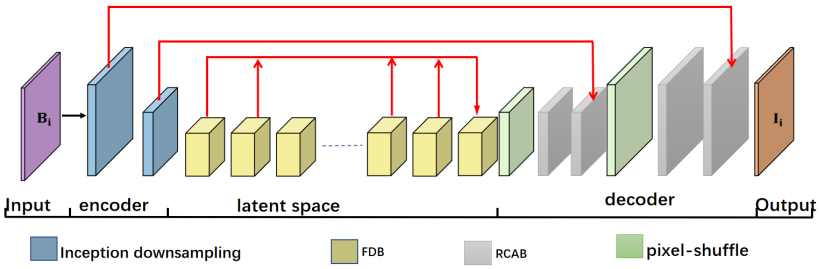


Figure 1. Overview the frequency disentanglement distillation image deblurring network (FDDN).

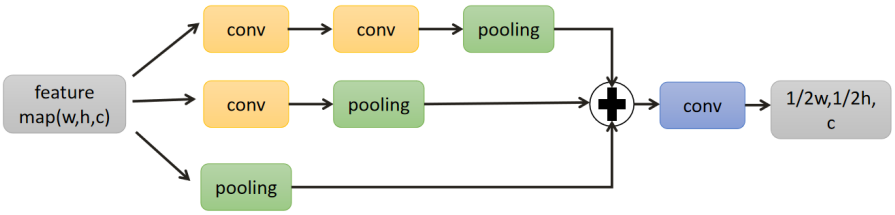


Figure 2. Inception down-sampling block.

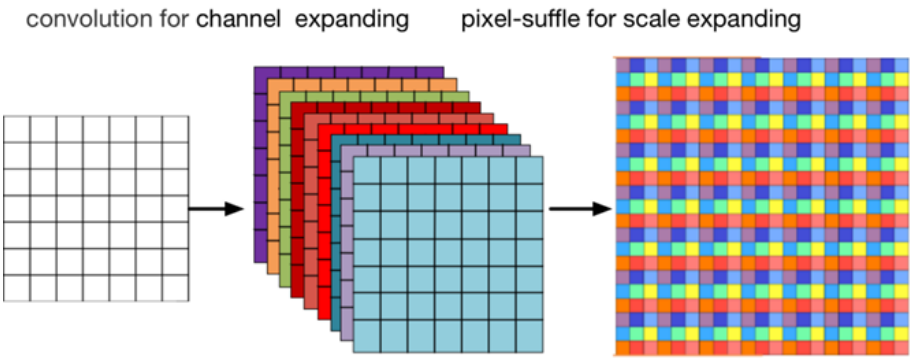


Figure 3. The architecture of pixel-shuffle block.

3.2. The Algorithm Frequency Split Block

Inspired by octconv [15], and in response to the needs of our network itself, we propose a frequency-based split module. As shown in Figure 4, Where the grey block representing the input feature map, red block representing high-frequency feature map, and blue block representing low-frequency. The frequency split block(FSB) is the component of the frequency distillation block. It is responsible for completing the channel split task during the distillation. Through the channel split, the useful part is retained, and the less useful part continues to be recursively distilled. Frequency split block splits the channels from the perspective of high and low frequencies, making the network more interpretable and efficient. The frequency split block still follows the design of octconv [15], which can carry out communication in intra-frequency, as well as inter-frequency. FSB is roughly divided into the following four steps, as shown in Table 1. Step 1, we determine the hyperparameter ratio. In order to meet the flexibility of the network, we can set different ratios according to the distribution of the dataset. The high frequency channel number is HChannel, and the low frequency channel number is LChannel. In the Step 2, the input feature map through Conv_croase2h and Conv_croase2l divided into two parts, feature to high and feature to low. This time, the high and low frequencies that are distinguished are only roughly divided. In the subsequent of the algorithm, the intra-frequency and the inter-frequency communication are are distinguished in more detail. Since the mostly redundant information in low-frequency, we perform down-sampling in the low-frequency information. Here, the down-sampling is done by a convolution operation with the stride of 2. The Step 3 is to complete the intra-frequency communication by convh2h and convl2l.

The Step 4 is inter-frequency communication. In the conversion process between high and low frequency, it will be accompanied by the transformation of the feature scale. In order to remove the redundancy of low-frequency information. Finally, through such a frequency dimension-based disentanglement method, the input features are generated under the operation of keeping the size unchanged, and two different features of high and low are generated. The parameters of convolution are shown in Table 2.

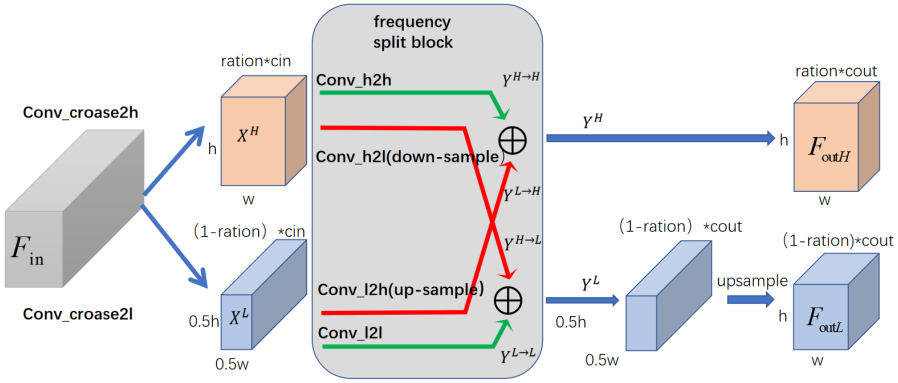


Figure 4. The construction of frequency split block(FSB). The gray block represent input feature maps of FSB. The blue represents low frequency channel and the red means high frequency channels.

Table 1. The algorithm frequency split block.

Input: $F_{in}(w, h, input), ratio, input\_channel, output\_channel$
Step1: Hchannel = ratio * output_channnel
Lchannel = (1-ratio) * output_channel
Step2: Feature to High = lucky_relu[Conv_croase2h(Feature in)]
Feature to Low = lucky_relu[Conv_croase2l(down-sampling(Feature in))]
Step3: H2h = Conv_h2h(feature to high)
L2l = Conv_l2l(feature to low))]
Step4: h2l = lucky_relu[Conv_h2l(down-sampling(feature to high))]
L2h = lucky_relu[Conv_l2h(up-sampling(feature to low))]
Output: $F_h = H2h + L2h$
$F_l = up - sampling(L2l + h2l)$

Table 2. The detail parameters of the convolution layer of FSB.

Conv_Name	Input_Channel	Output_Channel	Kernal-Size	Stride
Conv_croase2h	input_channel	Hchannel	3	1
Conv_croase2l	input_channel	Lchannel	3	1
Conv_h2h	Hchannel	Hchannel	1	1
Conv_h2l	Lchannel	Lchannel	1	1
Conv_l2l	Lchannel	Lchannel	1	1
Conv_l2h	Lchannel	Hchannel	1	1

3.3. Frequency Distillation Block (FDB)

Frequency distillation block is used in latent space. There are two reasons for this. The first is to improve the performance of the network. There are high-dimensional features in the latent space, and the feature size is generally small, which makes the amount of calculation less during the convolution operation. Second, our purpose is to distinguish feature information in the frequency dimension. High-dimensional features make it easier to distinguish between foreground semantic information and background information, as well as areas with rich features and flat gradients. This makes the high-dimensional features in latent space suitable for entangled operations. As show in Figure 5a, The traditional distillation block [28] is a progressive refinement module. It employs three channel split operations on the preceding features, which will produce two-part features. This one will be sent directly through the jump link mechanism in the feature fusion stage; this part's

channels are regarded as the useful information for restoring the image. The remaining part will be sent to the next recursive channel split operation. However, it simply divides the feature channel only according to a ratio. The detail of traditional distillation block [28] is show in Equations (2)–(5).

$$F_{distilled1},\ F_{coarse1} = Split_1(L_1(F_{in})) \tag{2}$$

$$F_{distilled2},\ F_{coarse2} = Split_2(L_2(F_{coarse1})) \tag{3}$$

$$F_{distilled3},\ F_{coarse3} = Split_3(L_3(F_{coarse2})) \tag{4}$$

$$F_{distilled4} = L_4(F_{coarse3})) \tag{5}$$

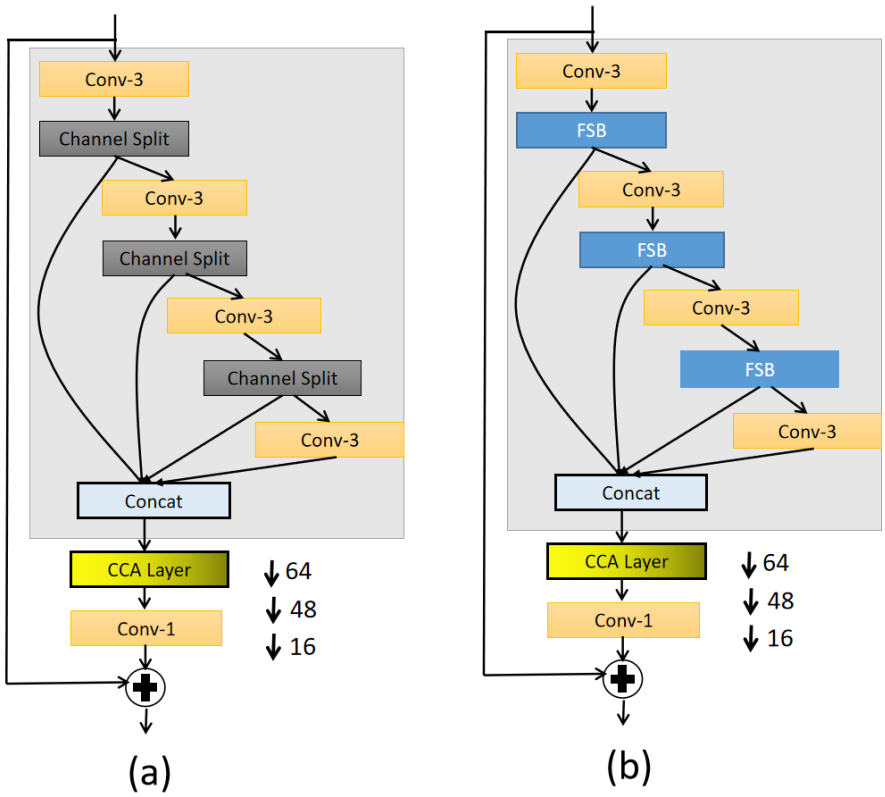


Figure 5. (**a**) represent tradition distillation block; (**b**) the disgn of FDB.

Therefore, in this article, we propose a frequency distillation block (FDB), as show in Figure 5b , which is regarded as a frequency-based disentanglement. FDB uses the frequency split block (FSB) of Section 3.1 mentioned above. $F_{low_n}(n = 1, 2, 3)$ is the low-frequency feature in the feature map, and $F_{high_n}(n = 1, 2, 3, 4)$ is the high-frequency feature. By keeping the high-frequency features directly, we think that the high-frequency features contain rich details such as the contours and edges of the foreground in the image. In addition, F_{low_n}, which is sent to the recursive distillation process again, is the low-frequency feature. We believe that in the low-frequency feature, the gradient changes slowly. The low-frequency mainly contains information such as the color, lighting, and blur features of the image. Of course, it does not completely rule out the existence of some useful details, so through the next distillation, purification is retained. The operation after this is to complete the operation of FDB through feature fusion with the saved $F_{high_n}(n = 1, 2, 3, 4)$. Subsequent operations merge the saved $F_{high_n}(n = 1, 2, 3, 4)$ to get the output feature F_{out}. The detail of FDB is show in Equations (6)–(11).

$$F_{high1},\ F_{low1} = FSB_1(L_1(F_{in})) \tag{6}$$

$$F_{high2},\ F_{low2} = FSB_2(L_2(F_{low1})) \tag{7}$$

$$F_{high3},\ F_{low3} = FSB_3(L_3(F_{low2})) \tag{8}$$

$$F_{high4} = L_4(F_{low3}) \tag{9}$$

$$F = CAT(F_{high4}, F_{high2}, F_{high3}, F_{high4}) \tag{10}$$

$$F_{out} = CCA(F) \tag{11}$$

3.4. Loss Function

3.4.1. Mse Loss

When training the generative adversarial network, it is necessary to compare the reconstructed image with ground truth by appropriate measurement. Usually, people use a pixel-by-pixel comparison loss function to measure the difference between the reconstructed image and the ground truth. However, using the pixel-by-pixel comparison loss function alone will produce artifacts. For example, consider two identical images that are offset from each other by one pixel; although they are very similar in perception, the results will be very different. In this case, the network will use the average of all possible solutions as the convergence value, which will cause artifacts. However, the pixel-by-pixel loss can still retain the detailed information of the picture to a certain extent. Therefore, we choose MSE as the pixel-by-pixel loss function, but we give it a relatively small weight. See Equation (12) for specific details.

$$L_{MSE} = \frac{1}{wh}\sum_{x=1}^{w}\sum_{y=1}^{h}((l_i)_{x,y} - G_{\theta G}(B_i)_{x,y})^2 \tag{12}$$

Among them, B_i represents the input blurred picture, $G_{\theta G}$ represents the generation network, and l_i is the standard clear image. The w and h are the length and width of the input/output image, respectively.

3.4.2. Perception Loss

At the same time, research [29] shows that perception loss can make the generation network improve the image quality. It maps the real picture and the generated picture to the feature map of the deep network and then calculates the least square method based on the feature map. This solves the disadvantage of pixel-by-pixel loss and performs pixel-by-pixel difference on the mapped feature map so that even if there is a certain degree of displacement, it will not have much impact. The process is shown in Equation (13), where w, h is the length and width of the feature map, and the parameters of the feature map are obtained by the VGG-16 network in the ReLU 3_3 layer.

$$L_{percep} = \frac{1}{wh}\sum_{x=1}^{w}\sum_{y=1}^{h}(\varphi(l_i)_{x,y} - \varphi(G_{\theta G}(B_i))_{x,y})^2 \tag{13}$$

The total loss function is shown in Equation (14).

$$L_{Total} = L_{MSE} + \lambda L_{Percep} \tag{14}$$

4. Experiments

4.1. Dataset

In order to prove the effectiveness of the frequency disentangle distillation image deblurring network (FDDN) more convincingly, and to avoid the situation that the network has excellent performance only on a specific dataset due to over-fitting. We will conduct comparative experiments on three different datasets.

GoPro [2] dataset uses GoPro Hero 4 camera to capture video sequences at 240 frames per second (fps). This dataset consists of 3214 pairs of blurry and sharp images with a resolution of 1280 $\times$ 720. Among them, 1111 pairs are used as the testset. Different from using the blur kernel to convolve on a sharp image to obtain a blurred image, GoPro [2]

follows the approximate camera imaging process during the image generation process in the blur and integrates consecutive frames within a certain exposure time to highlight the exposure time. The movement of the object inside is caused by the artifacts caused by the displacement, thereby generating a blurred image, rather than assuming a specific movement and designing a complex blur kernel. Therefore, there are only pairs of sharp/blurred image pairs in the dataset, and with no blur kernel. This kind of deblurring dataset without kernel estimation, compared with the traditional synthetic deblurring dataset with uniform blur kernel, is in the foreground, and the static background shows more realistic spatial blur changes.

HIDE [30] dataset is carefully constructed for human-aware image deblurring, covering a wide range of scenes, motions. HIDE dataset has 8422 sharp and blurry image pairs, extensively annotated with 65,784 human bounding boxes. For evaluation purposes, the images are split into separate training and test sets. Following random selection, we arrive at a unique split containing 6397 training and 2025 test images.

In this paper, we set a new Karate dataset in real scenes. It is difficult to get a sharp image completely corresponding to the blurred image after obtaining the blurred image. Even if certain conditions are deliberately created, slight deviations are unavoidable. Therefore, most of the benchmark datasets are obtained by the synthesis to obtain the paired images at this stage. There is no way to verify the ability of the algorithm to deblur the blurred image directly obtained in the real world. Therefore, we built a blurted dataset of the real scene. The main scene of the dataset is a karate match, and it is unpaired, with only blurred images and no corresponding ground truth.

4.2. Training Details

The experience environment parameters were as follows: Intel Core i5 9400F CPU@2.9GHz; memory: 32.00 GB; operating system: Ubuntu18.04; GPU: Nvidia RTX2080Ti. We obtained the following fixed parameters through repeated experiments and adjustments: ratio = 0.5; λ = 0.1. The training uses 2500 epochs. The learning rate is 0.001, and each iteration attenuates 0.0000001 after 500 epochs; the optimizer uses adam; it was trained on a RTX2080Ti for about 14 days. Since it is a fully convolutional network, images of any size can be accepted. Data enhancement options such as horizontal flip, quality compression, rotation, optical transformation, color change, cropping, hue saturation transformation, motion blur, median blur, snow scene and grayscale image conversion, are shown in Figure 6.

4.3. Quantitative and Qualitative Evaluation on Gopro Dataset

We evaluate the performance and efficiency of our model in the GoPro [2] dataset. We make comparisons with the state-of-the-art deblurring methods [2,16,17,27–29] in Pre-Processing, in terms of PSNR, SSIM, model size and inference time for images. The quantitative results are shown in Table 3. Visual comparisons are shown in Figure 7.

4.4. Quantitative and Qualitative Evaluation on Hide Dataset

To verify the validity of our method, we further evaluate our approach on the HIDE testing set [28]. In Table 4, we show a comparison with some of the methods. Visual comparisons are shown in Figure 8. From the data, we can see that our proposed method performs very well on this database.

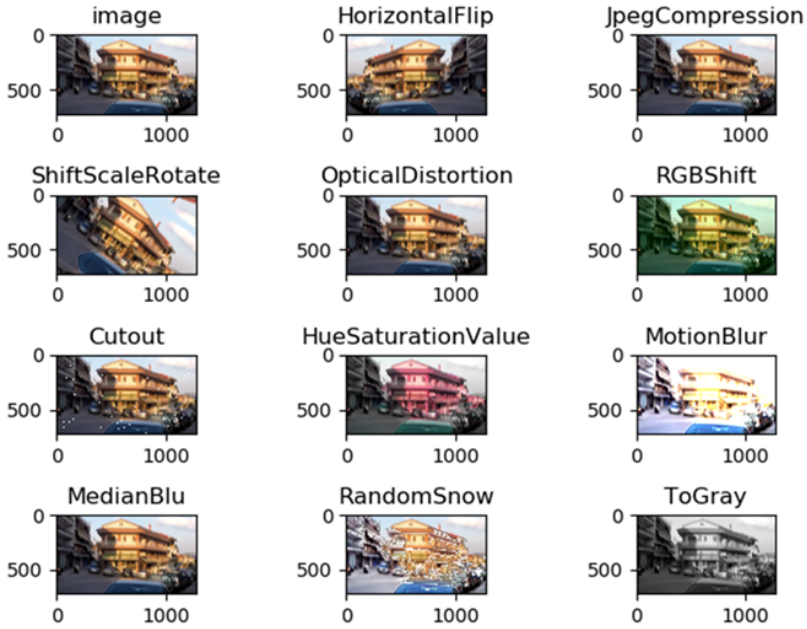


Figure 6. The visual represent of data enhancement.

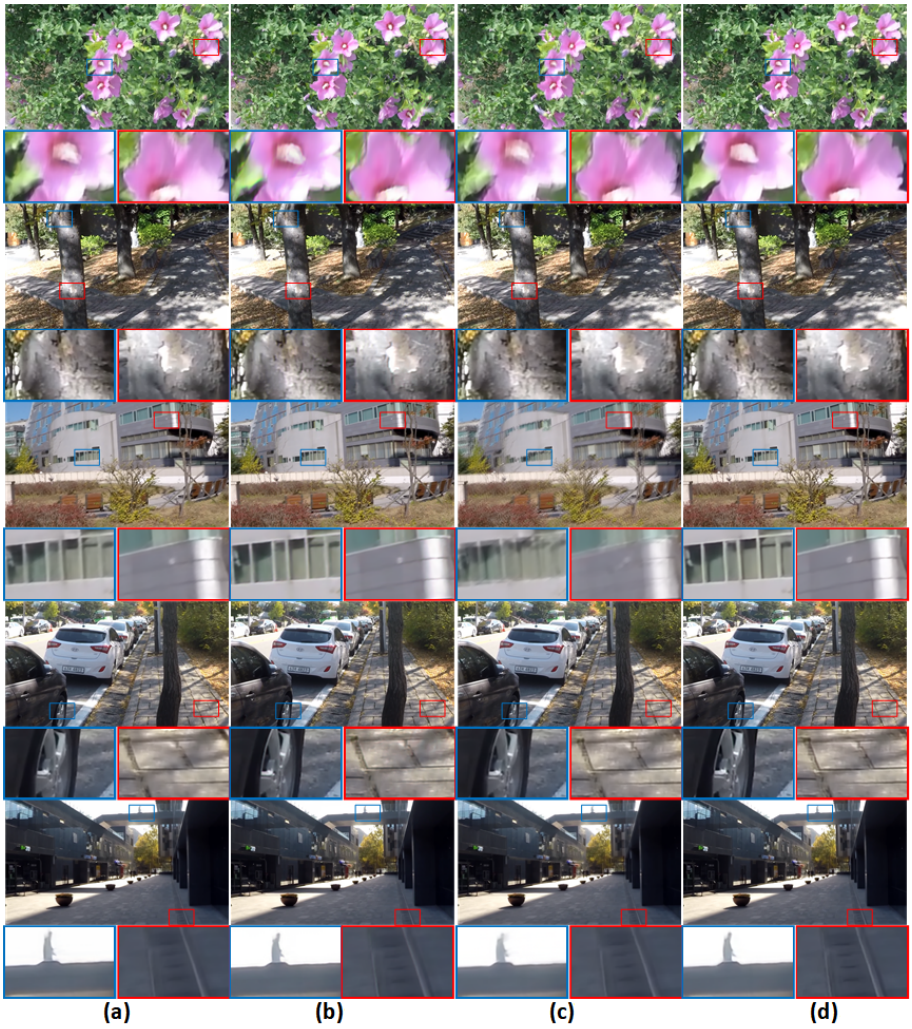


Figure 7. Visual comparison example on the GoPro dataset. (**a**) DMPHN [16]; (**b**) Gao et al. [17]; (**c**) Tao et al. [31]; (**d**) Ours.

Table 3. Performance and efficiency comparison on the GoPro dataset.

Methods	PSNR	SSIM	Model Size (MB)	Time (s)
DeepDeblur [2]	29.08	0.841	303.6	15
Zhang et al. [3]	29.19	0.9306	37.1	1.4
Gao et al. [17]	30.92	0.9421	2.84	1.6
DeblurGAN [5]	28.70	0.927	37.1	0.85
Tao et al. [31]	30.10	0.9323	33.6	1.6
DeblurGANv2 [9]	29.55	0.934	15	0.35
DMPHN [16]	30.21	0.9345	21.7	0.03
SIS [32]	30.28	0.912	36.54	0.303
Yuan et al. [33]	29.81	0.936	3.1	0.01
Pan et al. [20]	31.40	0.947	-	-
Wu et al. [21]	30.75	0.913	29.1	3.2
SharpGAN. [22]	29.62	0.897	-	0.17
Ours	**31.42**	**0.923**	**8.08**	**0.019**

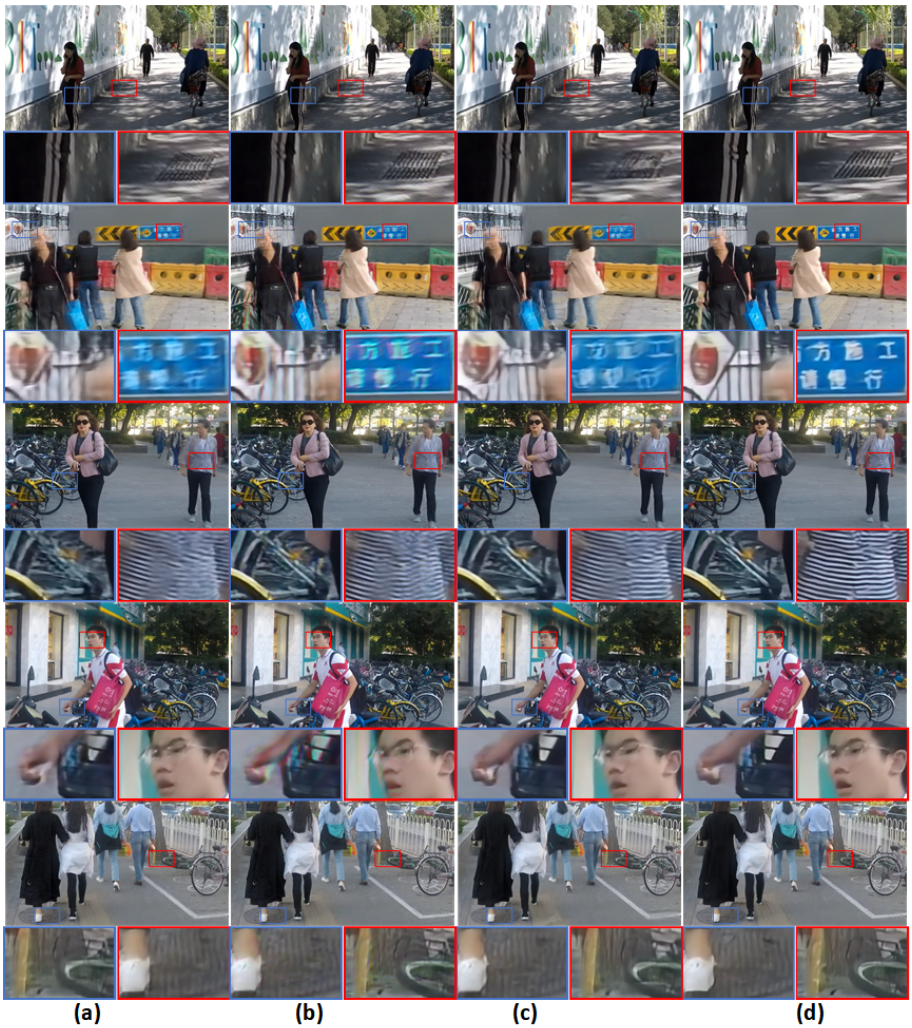


Figure 8. Visual comparison example on the HIDE dataset. (**a**) DMPHN [16]; (**b**) Tao et al. [31]; (**c**) Kupyn et al. [5]; (**d**) Ours.

Table 4. Performance comparison on the HIDE dataset.

Methods	Sun et al. [34]	DMPHN [16]	Nah et al. [2]	Tao et al. [31]	Kupyn et al. [5]	GCResNet [19]	FDDN (Ours)
PSNR	23.21	29.09	27.43	28.60	26.44	30.04	**30.07**
SSIM	0.797	0.930	0.902	0.928	0.890	0.924	**0.923**

4.5. Qualitative Evaluation of the Real-World Dataset

In this section, we have collected a set of datasets about karate competitions. The dataset is unpaired, with only blurred images and no ground. Therefore, PSNR and SSIM cannot be calculated without ground truth. Only in Figure 9 is the deblurring effect visualized.

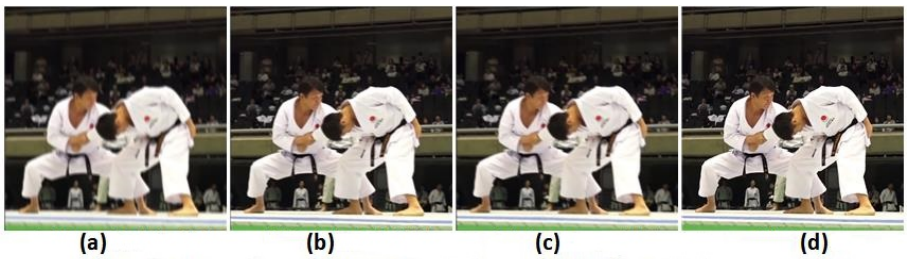


Figure 9. Visual comparison example on the karate dataset. (**a**) DeepDeblur [2]; (**b**) Deblur [5]; (**c**) Zhang et al. [3]; (**d**) Ours.

4.6. Ablation Study

In this part, we verify the effectiveness of the method proposed in this paper through 3 sets of comparative experiments. The verified modules are frequency split block, distillation block, and frequency distillation block. In the first set of ablation experiments, we replaced the distillation block with resnet [35]. In order to preserve the frequency split block, we added the frequency split block to the feature pixel level to achieve feature fusion. We found that the absence of the distillation block will also greatly affect the model effect. In the second set of ablation experiments, to eliminate the frequency split block, we used RFDB [27] to replace the frequency distillation block. The channel segmentation of RFDB [27] uses a single convolution operation. Because a simple structure is used to replace a complex structure, there will be a slight advantage in the model size. However, through experimental results, it is found that the deblurring effect of the model will be affected. In the third set of ablation experiments, we used resnet [35] instead of the complete frequency distillation block. It was found that the deblurring ability of the model dropped the most in all ablation experiments, indicating that no matter the distillation block, the frequency split block, or the frequency distillation block composed of them, they all played a key role in the model. The specific quantitative data can be seen in Table 5.

Table 5. Quantitative comparison of different ablations of our network on the GroPro dataset.

Distillation Block	Frequency Split Block	Frequency Distillation Block	PSNR	SSIM	Model Size (MB)
✗	✓	✓	29.65	0.892	9.05
✓	✗	✓	29.80	0.901	7.96
✓	✓	✗	**29.21**	**0.863**	**7.89**
✓	✓	✓	**31.42**	**0.923**	**8.08**

4.7. Analysis of the FDDN

Practical advantages: FDDN has achieved convincing results in the three parameters of PSNR, SSIM and model size, which means that this model has certain advantages in running speed and running effect. Due to the design of the FDB, the FDDN has a very deep model structure, which means every pixel of feature map have a very large receptive field. This property allows the image restoration process can make better use of the surrounding pixel information to restore the image details. The details can be see in Figure 8. In addition, according to Experiment 4.6, it can be seen that FDDN can not only recover image blur well on public datasets, but also generalize to specific application scenarios, such as motion blur in karate.

Disadvantages: FDDN cannot directly solve the entanglement of the blur and content information, but indirectly realizes the entanglement of blur information and content information through distillation operation in the two dimensions of high and low frequency. High-frequency information is retained as far as possible, and the redundancy of low-frequency information is eliminated.

5. Conclusions

Image deblurring is an important technical means to ensure the quality of the image. In this paper, we hope to realize the disentanglement of blur information and content information from the perspective of frequency. Therefore, we proposed the frequency disentanglement distillation image deblurring network (FDDN), which have three contribution: first, we proposed the frequency split block (FSB), which can distill high-frequency and low-frequency in defferent channels. Second, frequency distillation Block (FDB), which is a combination of frequency split block (FSB) and distillation block. FDB can be regarded as a frequency-based disentanglement. By keeping the high-frequency features directly and sending the low-frequency feature to the recursive distillation process, FDB distillate the useful feature step by step. Third, we perform extensive experiments on the tasks of motion deblurring using both synthetic datasets and real images and achieve an efficient result. We find that the FDDN have a good ability of generalization, and it can restore the details of blurry area effectively.

In the following work, we will further explore and improve the image restore ability of FDDN, which is not only used for image blurring, but also can be extended to image derain, super resolution, image inpainting and other joint tasks. In addition, the transformer mechanism will be introduced to further improve the quality of image restore. Finally, we will reduce the parameters of the model, so that FDDN can complete the real-time deblurring task.

Author Contributions: Conceptualization: J.G. and Y.L.; Funding acquisition: J.G.; Investigation: H.Z. and M.L.; Methodology: Y.L. and S.Y.; Project administration: D.X. and X.L.; Software: Y.L. and S.Y.; Visualization: T.L.; Writing—original draft: Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the research on image restoration algorithm based on regularization method (No. 10JJ3060).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code of FDDN can be available at https://github.com/yimingliu123/FDDN (accessed on 7 June 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tai, Y.W.; Tan, P.; Brown, M.S. Richardson-Lucy Deblurring for Scenes under a Projective Motion Path. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1603–1618. [PubMed]
2. Nah, S.; Kim, T.H.; Lee, K.M. Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. *arXiv* **2016**, arXiv:1612.02177.
3. Zhang, J.; Pan, J.; Ren, J.; Song, Y.; Yang, M.H. Dynamic Scene Deblurring Using Spatially Variant Recurrent Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
4. Aittala, M.; Durand, F. Burst Image Deblurring Using Permutation Invariant Convolutional Neural Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
5. Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; Matas, J. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
6. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Bing, X.; Bengio, Y. *Generative Adversarial Nets*; MIT Press: Cambridge, MA, USA, 2014.

7. Zoran, D.; Weiss, Y. From learning models of natural image patches to whole image restoration. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 479–486.
8. Dong, G.; Jie, Y.; Liu, L.; Zhang, Y.; Shi, Q. From Motion Blur to Motion Flow: A Deep Learning Solution for Removing Heterogeneous Motion Blur. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
9. Kupyn, O.; Martyniuk, T.; Wu, J.; Wang, Z. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 21–26 July 2017.
10. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
11. Chen, L.; Fang, F.; Wang, T.; Zhang, G. Blind Image Deblurring With Local Maximum Gradient Prior. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
12. Li, L.; Pan, J.; Lai, W.; Gao, C.; Sang, N.; Yang, M. Blind Image Deblurring via Deep Discriminative Priors. *Int. J. Comput. Vis.* **2019**, *127*, 1025–1043. [CrossRef]
13. Pan, J.; Sun, D.; Pfister, H.; Yang, M.H. Blind Image Deblurring Using Dark Channel Prior. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
14. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.; Wong, W.; Woo, W. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.
15. Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; Feng, J. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks With Octave Convolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2020.
16. Zhang, H.; Dai, Y.; Li, H.; Koniusz, P. Deep Stacked Hierarchical Multi-patch Network for Image Deblurring. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
17. Gao, H.; Tao, X.; Shen, X.; Jia, J. Dynamic Scene Deblurring with Parameter Selective Sharing and Nested Skip Connections. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
18. Lu, B.; Chen, J.C.; Chellappa, R. Unsupervised Domain-Specific Deblurring via Disentangled Representations. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
19. Xu, B.; Yin, H. Graph Convolutional Networks in Feature Space for Image Deblurring and Super-resolution. *arXiv* **2021**, arXiv:2105.10465.
20. Pan, Z.; Lv, Q.; Tan, Z. A Two-Stage Network for Image Deblurring. *IEEE Access* **2021**, *9*, 76707–76715. [CrossRef]
21. Wu, Y.; Qian, P.; Zhang, X. Two-Level Wavelet-Based Convolutional Neural Network for Image Deblurring. *IEEE Access* **2021**, *9*, 45853–45863. [CrossRef]
22. Feng, H.; Guo, J.; Xu, H.; Ge, S.S. SharpGAN: Dynamic Scene Deblurring Method for Smart Ship Based on Receptive Field Block and Generative Adversarial Networks. *Sensors* **2021**, *21*, 3641. [CrossRef] [PubMed]
23. Wang, H.; Yue, Z.; Zhao, Q.; Meng, D. A Deep Variational Bayesian Framework for Blind Image Deblurring. *arXiv* **2021**, arXiv:2106.02884.
24. Zamir, S.W.; Arora, A.; Khan, S.H.; Hayat, M.; Khan, F.S.; Yang, M.; Shao, L. Multi-Stage Progressive Image Restoration. *arXiv* **2021**, arXiv:2102.02808.
25. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
26. Liu, J.; Tang, J.; Wu, G. Residual Feature Distillation Network for Lightweight Image Super-Resolution. In Proceedings of the Computer Vision—ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Bartoli, A., Fusiello, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12537, pp. 41–55.
27. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
28. Hui, Z.; Gao, X.; Yang, Y.; Wang, X. Lightweight Image Super-Resolution with Information Multi-distillation Network. In Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, 21–25 October 2019; Amsaleg, L., Huet, B., Larson, M.A., Gravier, G., Hung, H., Ngo, C., Ooi, W.T., Eds.; ACM: New York, NY, USA, 2019; pp. 2024–2032.
29. Cai, J.; Zuo, W.; Zhang, L. Dark and Bright Channel Prior Embedded Network for Dynamic Scene Deblurring. *IEEE Trans. Image Process.* **2020**, *29*, 6885–6897. [CrossRef]
30. Shen, Z.; Wang, W.; Lu, X.; Shen, J.; Ling, H.; Xu, T.; Shao, L. Human-Aware Motion Deblurring. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5571–5580.
31. Tao, X.; Gao, H.; Wang, Y.; Shen, X.; Wang, J.; Jia, J. Scale-recurrent Network for Deep Image Deblurring. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
32. Liu, Y.; Luo, Y.; Huang, W.; Qiao, Y.; Luo, D. Semantic Information Supplementary Pyramid Network for Dynamic Scene Deblurring. *IEEE Access* **2020**, *8*, 188587–188599. [CrossRef]

33. Yuan, Y.; Su, W.; Ma, D. Efficient Dynamic Scene Deblurring Using Spatially Variant Deconvolution Network With Optical Flow Guided Training. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
34. Sun, J.; Cao, W.; Xu, Z.; Ponce, J. Learning a Convolutional Neural Network for Non-uniform Motion Blur Removal. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778; [CrossRef]

dMELODIES: A MUSIC DATASET FOR DISENTANGLEMENT LEARNING

Ashis Pati **Siddharth Gururani** **Alexander Lerch**
Center for Music Technology, Georgia Institute of Technology, USA
ashis.pati@gatech.edu, siddgururani@gatech.edu, alexander.lerch@gatech.edu

ABSTRACT

Representation learning focused on disentangling the underlying factors of variation in given data has become an important area of research in machine learning. However, most of the studies in this area have relied on datasets from the computer vision domain and thus, have not been readily extended to music. In this paper, we present a new symbolic music dataset that will help researchers working on disentanglement problems demonstrate the efficacy of their algorithms on diverse domains. This will also provide a means for evaluating algorithms specifically designed for music. To this end, we create a dataset comprising of 2-bar monophonic melodies where each melody is the result of a unique combination of nine latent factors that span ordinal, categorical, and binary types. The dataset is large enough ($\approx$ 1.3 million data points) to train and test deep networks for disentanglement learning. In addition, we present benchmarking experiments using popular unsupervised disentanglement algorithms on this dataset and compare the results with those obtained on an image-based dataset.

1. INTRODUCTION

Representation learning deals with extracting the underlying factors of variation in a given observation [1]. Learning compact and *disentangled* representations (see Figure 1 for an illustration) from given data, where important factors of variation are clearly separated, is considered useful for generative modeling and for improving performance on downstream tasks (such as speech recognition, speech synthesis, vision and language generation [2–4]). Disentangled representations allow a greater degree of interpretability and controllability, especially for content generation, be it language, speech, or music. In the context of Music Information Retrieval (MIR) and generative music models, learning some form of disentangled representation has been the central idea for a wide variety of tasks such as genre transfer [5], rhythm transfer [6, 7], timbre synthesis [8], instrument rearrangement [9], manipulating musical attributes [10, 11], and learning music similarity [12].

Consequently, there exists a large body of research in the machine learning community focused on developing algorithms for learning disentangled representations. These span unsupervised [13–16], semi-supervised [17–19] and supervised [10, 20–22] methods. However, a vast majority of these algorithms are designed, developed, tested, and evaluated using data from the image or computer vision domain. The availability of standard image-based datasets such as dSprites [23], 3D-Shapes [24], and 3D-Chairs [25] among others has fostered disentanglement studies in vision. Additionally, having well-defined factors of variation (for instance, size and orientation in dSprites [23], pitch and elevation in Cars3D [26]) has allowed systematic studies and easy comparison of different algorithms. However, this restricted focus on a single domain raises concerns about the generalization of these methods [27] and prevents easy adoption into other domains such as music.

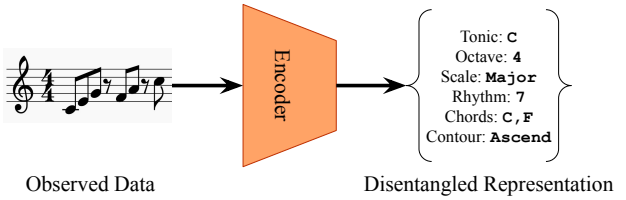


Figure 1: Disentanglement example where a high dimensional observed data is disentangled into a low dimensional representation comprising of semantically meaningful factors of variation.

Research on disentanglement learning in music has often been application-oriented with researchers using their own problem-specific datasets. The factors of variation have also been chosen accordingly. To the best of our knowledge, there is no standard dataset for disentanglement learning in music. This has prevented systematic research on understanding disentanglement in the context of music.

In this paper, we introduce *dMelodies*, a new dataset of monophonic melodies, specifically intended for disentanglement studies. The dataset is created algorithmically and is based on a simple and yet diverse set of independent latent factors spanning ordinal, categorical and binary attributes. The full dataset contains $\approx$ 1.3 million data points which matches the scale of image datasets and should be sufficient to train deep networks. We consider this dataset as the primary contribution of this paper. In addition, we also conduct benchmarking experiments using three popular unsupervised methods for disentanglement learning and present a comparison of the results with the dSprites dataset [23]. Our experiments show that disentanglement

learning methods do not directly translate between the image and music domains and having a music-focused dataset will be extremely useful to ascertain the generalizability of such methods. The dataset is available online [1] along with the code to reproduce our benchmarking experiments. [2]

2. MOTIVATION

In representation learning, given an observation $\mathbf{x}$, the task is to learn a representation $r(\mathbf{x})$ which "makes it easier to extract useful information when building classifiers or other predictors" [1]. The fundamental assumption is that any high-dimensional observation $\mathbf{x} \in \mathcal{X}$ (where $\mathcal{X}$ is the data-space) can be decomposed into a semantically meaningful low dimensional latent variable $\mathbf{z} \in \mathcal{Z}$ (where $\mathcal{Z}$ is referred to as the latent space). Given a large number of observations in $\mathcal{X}$, the task of disentanglement learning is to estimate this low dimensional latent space $\mathcal{Z}$ by separating out the distinct factors of variation [1]. An ideal disentanglement method ensures that changes to a single underlying factor of variation in the data changes only a single factor in its representation [27]. From a generative modeling perspective, it is also important to learn the mapping from $\mathcal{Z}$ to $\mathcal{X}$ to enable better control over the generative process.

2.1 Lack of diversity in disentanglement learning

Most state-of-the-art methods for unsupervised disentanglement learning are based on the Variational Auto-Encoder (VAE) [28] framework. The key idea behind these methods is that factorizing the latent representation to have an aggregated posterior should lead to better disentanglement [27]. This is achieved using different means, e.g., imposing constraints on the information capacity of the latent space [13, 29, 30], maximizing the mutual information between a subset of the latent code and the observations [31], and maximizing the independence between the latent variables [14, 15]. However, unsupervised methods for disentanglement learning are sensitive to inductive biases (such network architectures, hyperparameters, and random seeds) and consequently there is a need to properly evaluate such methods by using datasets from diverse domains [27].

Apart from unsupervised methods for disentanglement learning, there has also been some research on semi-supervised [18, 19] and supervised [20, 21, 32, 33] learning techniques to manipulate specific attributes in the context of generative models. In these paradigms, a labeled loss is used in addition to the unsupervised loss. Available labels can be utilized in various ways. They can help with disentangling known factors (e.g., digit class in MNIST) from latent factors (e.g., handwriting style) [34], or supervising specific latent dimensions to map to specific attributes [10]. However, most of these approaches are evaluated using image domain datasets.

Tremendous interest from the machine learning community has led to the creation of benchmarking datasets (albeit image-based) specifically targeted towards disentanglement learning such as dSprites [23], 3D-Shapes [24], 3D-chairs [25], MPI3D [35], most of which are artificially generated and have simple factors of variation. While one can argue that artificial datasets do not reflect real-world scenarios, the relative simplicity of these datasets is often desirable since they enable rapid prototyping.

2.2 Lack of consistency in music-based studies

Representation learning has also been explored in the field of MIR. Much like images, learning better representations has been shown to work well for MIR tasks such as composer classification [36, 37], music tagging [38], and audio-to-score alignment [39]. The idea of disentanglement has been particularly gaining traction in the context of interactive music generation models [5, 6, 11, 33]. Disentangling semantically meaningful factors can significantly improve the usefulness of music generation tools. Many researchers have independently tried to tackle the problem of disentanglement in the context of symbolic music by using different musically meaningful attributes such as genre [5], note density [10], rhythm [6], and timbre [8]. However, these methods and techniques have all been evaluated using different datasets which makes a direct comparison impossible. Part of the reason behind this lack of consistency is the difference in the problems that these methods were looking to address. However, the availability of a common dataset allowing researchers to easily compare algorithms and test their hypotheses will surely aid systematic research.

3. dMELODIES DATASET

The primary objective of this work is to create a simple dataset for music disentanglement that can alleviate some of the shortcomings mentioned in Section 2: first, researchers interested in disentanglement will have access to more diverse data to evaluate their methods, and second, research on music disentanglement will have the means for conducting systematic, comparable evaluation. This section describes the design choices and the methodology used for creating the proposed *dMelodies* dataset.

While core MIR tasks such as music transcription, or tagging focus more on analysis of audio signals, research on generative models for music has focused more on the symbolic domain. Considering most of the interest in disentanglement learning stems from research on generative models, we decided to create this dataset using symbolic music representations.

3.1 Design Principles

To enable objective evaluation of disentanglement algorithms, one needs to either know the ground-truth values of the underlying factors of variation for each data point, or be able to synthesize the data points based on the attribute values. The dSprites dataset [23], for instance, consists of single images of different 2-dimensional shapes with simple attributes specifying the position, scale and orientation of these shapes against a black background. The design of our

[1] https://github.com/ashispati/dmelodies_dataset
[2] https://github.com/ashispati/dmelodies_benchmarking

dataset is loosely based on the dSprites dataset. The following principles were used to finalize other design choices:

(a) The dataset should have a simple construction with homogenous data points and intuitive factors of variation. It should allow for easy differentiation between data points and have clearly distinguishable latent factors.
(b) The factors of variation should be independent, i.e., changing any one factor should not cause changes to other factors. While this is not always true for real-world data, it enables consistent objective evaluation.
(c) There should be a clear one-to-one mapping between the latent factors and the individual data points. In other words, each unique combination of the factors should result in a unique data point.
(d) The factors of variation should be diverse. In addition, it would be ideal to have the factors span different types such as discrete, ordinal, categorical and binary.
(e) Finally, the different combinations of factors should result in a dataset large enough to train deep neural networks. Based on size of the different image-based datasets [23,40], we would require a dataset of the order of at least a few hundred thousand data points.

3.2 Dataset Construction

Considering the design principles outlined above, we decided to focus on monophonic pitch sequences. While there are other options such as polyphonic or multi-instrumental music, the choice of monophonic melodies was to ensure simplicity. Monophonic melodies are a simple form of music uniquely defined by the pitch and duration of their note sequences. The pitches are typically based on the key or scale in which the melody is being played and the rhythm is defined by the onset positions of the notes.

Since the set of all possible monophonic melodies is very large and heterogeneous, the following additional constraints were imposed on the melody in order to enforce homogeneity and satisfy the other design principles:

(a) Each melody is based on a scale selected from a finite set of allowed scales. This choice of scale also serves as one of the factors of variation. The melody will also be uniquely defined by the pitch class of the tonic (root pitch) and the octave number.
(b) In order to constrain the space of all possible pitch patterns within a scale, we restrict each melody to be an arpeggio over the standard I-IV-V-I cadence chord pattern. Consequently, each melody consists of 12 notes (3 notes for each of the 4 chords).
(c) In order to vary the pitch patterns, the direction of arpeggiation of each chord, i.e. up or down, is used as a latent factor. This choice adds a few binary factors of variation to the dataset.
(d) The melodies are fixed to 2-bar sequences with 8th note as the minimum note duration. This makes the dataset uniform in terms of sequence lengths of the data points and also helps reduce the complexity of the sequences. 2-bar sequences have been used in other music generation studies as well [10, 41]. We use a tokenized data representation such that each melody is a sequence of length 16.
(e) If we consider the space of all possible unique rhythms, the number of options will explode to $\binom{16}{12}$ which will be significantly larger than other factors of variation. Hence, we choose to break the latent factor for rhythm into 2 independent factors: rhythm for bar 1 and bar 2.
(f) The rhythm of a melody is based on the metrical onset position of the notes [42]. Consequently, rhythm is dependent on the number of notes. In order to keep rhythm independent from other factors, we constrain each bar to have 6 notes (play 2 chords) thereby obtaining $\binom{8}{6}$ options for each bar.

Factor	# Options	Notes
Tonic	12	C, C#, D, through B
Octave	3	Octave 4, 5 and 6
Scale	3	major, harmonic minor, and blues
Rhythm Bar 1	28	$\binom{8}{6}$, based on onset locations of 6 notes
Rhythm Bar 2	28	$\binom{8}{6}$, based on onset locations of 6 notes
Arp Chord 1	2	up/down, for Chord 1
Arp Chord 2	2	up/down, for Chord 2
Arp Chord 3	2	up/down, for Chord 3
Arp Chord 4	2	up/down, for Chord 4

Table 1: Table showing the different factors of variation for the dMelodies dataset. Since all factors of variation are independent, the total dataset contains 1,354,752 unique melodies.

Based on the above design choices, the dMelodies dataset consists of 2-bar monophonic melodies with 9 factors of variations listed in Table 1. The factors of variation were chosen to satisfy the design principles listed in Section 3.1. For instance, while melodic transformations such as repetition, inversion, retrograde would have made more musical sense, they did not allow creation of a large-enough dataset with independent factors of variation. The resulting dataset thus contains simple melodies which do not adequately reflect real-world musical data. A side-effect of this choice of factors is that some of them (such as arpeggiation direction and rhythm) affect only a specific part of the data. Since each unique combination of these factors results in a unique data point we get 1,354,752 unique melodies. Figure 2 shows one such melody from the dataset and its corresponding latent factors. The dataset is generated using the *music21* [43] python package.

4. BENCHMARKING EXPERIMENTS

In this section, we present benchmarking experiments to demonstrate the performance of some of the existing unsupervised disentanglement algorithms on the proposed dMelodies dataset and contrast the results with those obtained on the image-based dSprites dataset.

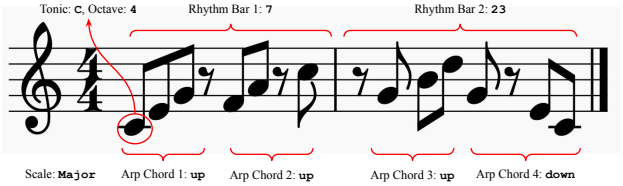


Figure 2: Example of a sample melody from the dMelodies dataset. Also shown are the values of the different latent factors. For rhythm latent factors, the shown value corresponds to the index from the rhythm dictionary.

4.1 Experimental Setup

We consider 3 different disentanglement learning methods: β-VAE [13], Annealed-VAE [29], and FactorVAE [15]. All these methods are based on different regularization terms applied to the VAE loss function.

4.1.1 Data Representation

We use a tokenized data representation [44] with the 8th-note as the smallest note duration. Each 8th note position is encoded with a token corresponding to the note name which starts on that position. A special continuation symbol ('__') is used which denotes that the previous note is held. A special token is used for rest.

4.1.2 Model Architectures

Two different VAE architectures are chosen to conduct these experiments. The first architecture (dMelodies-CNN) is based on Convolutional Neural Networks (CNNs) and is similar to those used for several image-based VAEs, except that we use 1-D convolutions. The second architecture (dMelodies-RNN) is based on a hierarchical recurrent model [41, 45]. Details of the model architectures are provided in the supplementary material.

4.1.3 Hyperparameters

Each learning method has its own regularizing hyperparameter. For β-VAE, we use three different values of $\beta \in \{0.2, 1.0, 4.0\}$. This choice is loosely based on the notion of normalized-β [13]. In addition, we force the KL-regularization only when the KL-divergence exceeds a fixed threshold $\tau = 50$ [41, 46]. For Annealed-VAE, we fix $\gamma = 1.0$ and use three different values of capacity, $C \in \{25.0, 50.0, 75.0\}$. For FactorVAE, we use the Annealed-VAE loss function with a fixed capacity ($C = 50$), and choose three different values for $\gamma \in \{1, 10, 50\}$.

4.1.4 Training Specifications

For each of the above methods, model, and hyperparameter combination, we train 3 models with different random seeds. To ensure consistency across training, all models are trained with a batch-size of 512 for 100 epochs. The ADAM optimizer [47] is used with a fixed learning rate of 1e$-$4, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\text{e}{-8}$. For β-VAE and Annealed-VAE, we use 10 warm-up epochs where $\beta = 0.0$. After warm-up, the regularization hyperparameter (β for β-VAE and C for Annealed-VAE) is annealed exponentially from 0.0 to their target values over 100000 iterations. For FactorVAE, we stick to the original implementation and do not anneal any of the parameters in the loss function. The VAE optimizer is the same as mentioned earlier. The FactorVAE discriminator is optimized using ADAM with a fixed learning rate of 1e$-$4, $\beta_1 = 0.8$, $\beta_2 = 0.9$, and $\epsilon = 1\text{e}{-8}$. We found that utilizing the original hyperparameters [15] for this optimizer led to unstable training on dMelodies.

For comparison with dSprites, we present the results for all the three methods using a CNN-based VAE architecture. The set of hyperparameters and other training configurations were kept the same for the dSprites dataset, except for the FactorVAE where we use the originally proposed loss function and discriminator optimizer hyperparameters, as the model does not converge otherwise.

4.1.5 Disentanglement Metrics

The following objective metrics for measuring disentanglement are used: (a) *Mutual Information Gap (MIG)* [14], which measures the difference of mutual information between a given latent factor and the top two dimensions of the latent space which share maximum mutual information with the factor, (b) *Modularity* [48], which measures if each dimension of the latent space depends on only one latent factor, and (c) *Separated Attribute Predictability (SAP)* [16], which measures the difference in the prediction error of the two most predictive dimensions of the latent space for a given factor. For each metric, the mean across all latent factors is used for aggregation. For consistency, standard implementations of the different metrics are used [27].

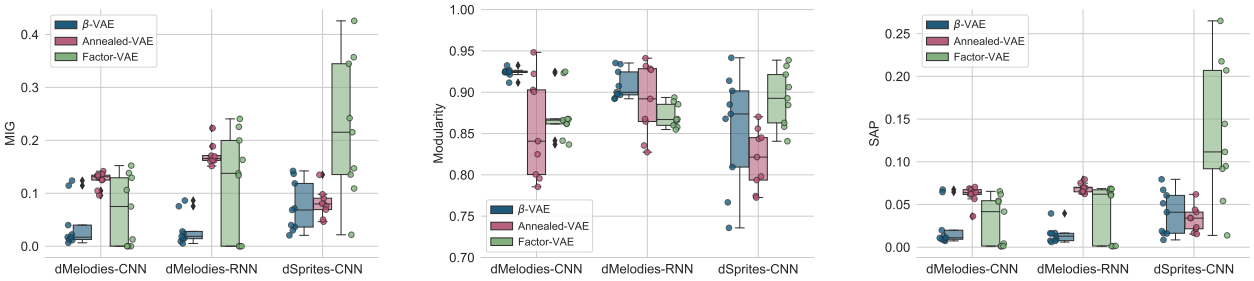
4.2 Experimental Results

4.2.1 Disentanglement

In this experiment, we present the comparative disentanglement performance of the different methods on dMelodies. The result for each method is aggregated across the different hyperparameters and random seeds. Figure 3 shows the results for all three disentanglement metrics. We group the trained models based on the architecture. The results for the dSprites dataset are also shown for comparison.

First, we compare the performance of different methods on dMelodies. Annealed-VAE shows better performance for MIG and SAP. These metrics indicate the ability of a method to ensure that each factor of variation is mapped to a single latent dimension. The performance in terms of Modularity is similar across the different methods. High Modularity indicates that each dimension of the latent space maps to only a single factor of variation. For dSprites, FactorVAE seems to be best method overall across metrics. However, the high variance in the results shows that choice of random seeds and hyperparameters is probably more important than the disentanglement method itself. This is in line with observations in previous studies [27].

Second, we observe no significant impact of model architecture on the disentanglement performance. For both the CNN and the hierarchical RNN-based VAE, the performance of all the different methods on dMelodies is



(a) *MIG* (b) *Modularity* (c) *SAP Score*

Figure 3: Overall disentanglement performance (higher is better) of different methods on the dMelodies and dSprites datasets. Individual points denote results for different hyperparameter and random seed combinations. Please refer to supplementary material Sec.2.1 for the best hyperparameter settings.

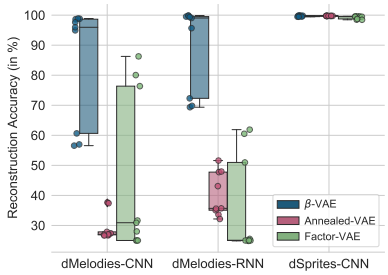


Figure 4: Overall reconstruction accuracies (higher is better) of the different methods on the dMelodies and dSprites datasets. Individual points denote results for different hyperparameter and random seed combinations.

comparable. This might be due to the relatively short sequence lengths used in dMelodies which do not fully utilize the capabilities of the hierarchical-RNN architecture (which has been shown to work well in learning long-term dependencies [41]). On the positive side, this indicates that the dMelodies dataset might be agnostic to the VAE-architecture.

Finally, we compare differences in the performance between the two datasets. In terms of MIG and SAP, the performance for dSprites is slightly better (especially for Factor-VAE), while for Modularity, performance across both datasets is comparable. However, once again, the differences are not significant. Looking at the disentanglement metrics alone, one might be tempted to conclude that the different methods are domain invariant. However, as the next experiments will show, there are significant differences.

4.2.2 *Reconstruction Fidelity*

From a generative modeling standpoint, it is important that along with better disentanglement performance we also retain good reconstruction fidelity. This is measured using the reconstruction accuracy shown in Figure 4. It is clear that all three methods fail to achieve a consistently good reconstruction accuracy on dMelodies. β-VAE gets an accuracy $\geq 90\%$ for some hyperparameter values (more on this in Section 4.2.3). However, both Annealed-VAE and Factor-VAE struggle to cross a median-accuracy of 40% (which would be unusable from a generative modeling perspective). The performance of the hierarchical RNN-based VAE is slightly better than the CNN-based architecture. In comparison, for dSprites, all three methods are able to consistently achieve better reconstruction accuracies.

4.2.3 *Sensitivity to Hyperparameters*

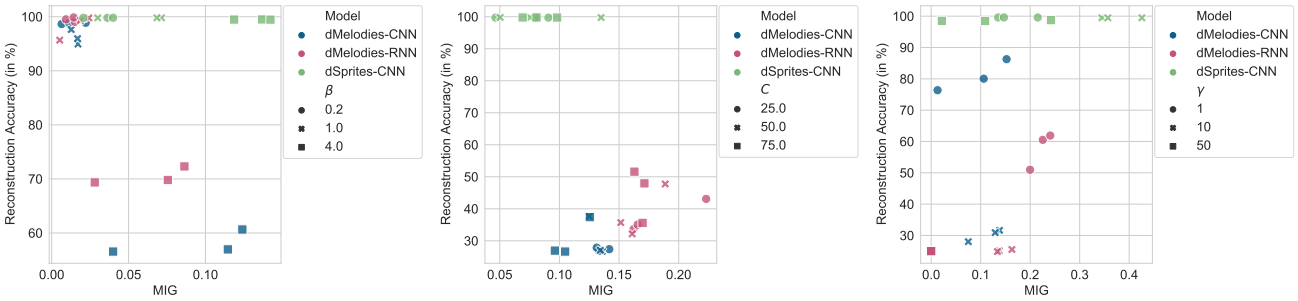
The previous experiments presented aggregated results over the different hyperparameter values for each method. Next, we take a closer look at the individual impact of those hyperparameters, i.e., the effect of changing the hyperparameters on the disentanglement performance (MIG) and the reconstruction accuracy. Figure 5 shows this in the form of scatter plots. The ideal models should lie on the top right corner of the plots (with high values of both reconstruction accuracy and MIG).

Models trained on dMelodies are very sensitive to hyperparameter adjustments. This is especially true for reconstruction accuracy. For instance, increasing β for the β-VAE model improves MIG but severely reduces reconstruction performance. For Annealed-VAE and Factor-VAE there is a wider spread in the scatter plots. For Annealed-VAE, having a high capacity C seems to marginally improve reconstruction (especially for the recurrent VAE). For FactorVAE, increasing γ leads to a drop in both disentanglement and reconstruction.

Contrast this with the scatter plots for dSprites. For all three methods, the hyperparameters seem to only significantly affect the disentanglement performance. For instance, increasing β and γ (for β-VAE and FactorVAE, respectively) result in clear improvement in MIG. More importantly, however, there is no adverse impact on the reconstruction accuracy.

4.2.4 *Factor-wise Disentanglement*

We also looked at how the individual factors of variation are disentangled. We consider the β-VAE model for this since it has the highest reconstruction accuracy. Figure 6



(a) β-VAE: Varying β (b) Annealed-VAE: Varying C (c) Factor-VAE: Varying γ

Figure 5: Effect of the hyperparameters on the different disentanglement methods. Overall, for improving disentanglement on dMelodies results in severe drop in reconstruction accuracy. The dSprites dataset does not suffer from this drawback.

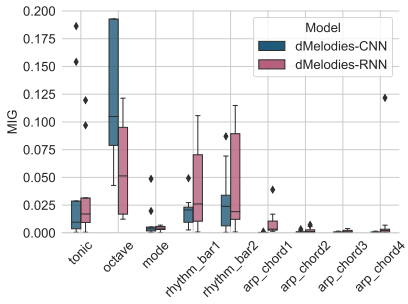


Figure 6: Factor-wise MIG for the β-VAE method.

shows the factor-wise *MIG* for both the CNN and RNN-based models. Factors corresponding to octave and rhythm are disentangled better. This is consistent with some recent research on disentangling rhythm [6, 7]. In contrast, the factors corresponding to the arpeggiation direction perform the worst. This might be due to their binary type. Similar analysis for the dSprites dataset reveals better disentanglement for the scale and position based factors. Additional results are provided in the supplementary material.

5. DISCUSSION

As mentioned in Section 2, disentanglement techniques have been shown to be sensitive to the choice of hyperparameters and random seeds [27]. The results obtained in our benchmarking experiments in the previous section using dMelodies seem to ascertain this even further. We find that methods which work well for image-based datasets do not extend directly to the music domain. When moving between domains, not only do we have to tune hyperparameters separately, but the model behavior may vary significantly when hyperparameters are changed. For instance, reconstruction fidelity is hardly effected by hyperparameter choice in the case of dSprites while for dMelodies it varies significantly. While sensitivity to hyperparameters is expected in neural networks, this is also one of the main reasons for evaluating methods on more than one dataset, preferably from multiple domains.

Some aspects of the dataset design, especially the nature of the factors of variation, might have affected our experimental results. While the factors of variation in dSprites are continuous (except the shape attribute), those for dMelodies span different data-types (categorical, ordinal and binary). This might make other types of models (such as VQ-VAEs [49]) more suitable. Another consideration is that some factors of variation (such as the arpeggiation direction and rhythm) effect only a part of the data. However, the effect of this on the disentanglement performance needs further investigation since we get good performance for rhythm but poor performance for arpeggiation direction.

Unsupervised methods for disentanglement learning have their own limitations and some degree of supervision might actually be essential [27]. It is still unclear if it is possible to develop general domain-invariant disentanglement methods. Consequently, supervised and semi-supervised methods have been garnering more attention [10, 11, 19, 34]. The dMelodies dataset can also be used to explore such methods for music-based tasks. There has been some work recently in disentangling musical attributes such as rhythm and melodic contours which are considered important from an interactive music generation perspective [6,11,50]. Apart from the designed latent factors of variation, other low-level musical attributes such as rhythmic complexity and contours can also be computationally extracted using this dataset to meet task-specific requirements.

6. CONCLUSION

This paper addresses the need for more diverse modes of data for studying disentangled representation learning by introducing a new music dataset for the task. The *dMelodies* dataset comprises of more than 1 million data points of 2-bar melodies. The dataset is constructed based on fixed rules that maintain independence between different factors of variation, thus enabling researchers to use it for studying disentanglement learning. Benchmarking experiments conducted using popular disentanglement learning methods show that existing methods do not achieve performance comparable to those obtained on an analogous image-based dataset. This showcases the need for further research on domain-invariant algorithms for disentanglement learning.

7. ACKNOWLEDGMENT

The authors would like to thank Nvidia Corporation for their donation of a Titan V awarded as part of the GPU (Graphics Processing Unit) grant program which was used for running several experiments pertaining to this research.

8. REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 2013.

[2] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, Long Beach, California, USA, 2017.

[3] W. Hsu, Y. Zhang, R. J. Weiss, Y. Chung, Y. Wang, Y. Wu, and J. R. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019*, Brighton, United Kingdom, 2019.

[4] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding," in *Advances in Neural Information Processing Systems 31 (NeurIPS)*, Montréal, Canada, 2018.

[5] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, "MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer," in *Proc. of 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.

[6] R. Yang, D. Wang, Z. Wang, T. Chen, J. Jiang, and G. Xia, "Deep music analogy via latent representation disentanglement," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[7] J. Jiang, G. G. Xia, D. B. Carlton, C. N. Anderson, and R. H. Miyakawa, "Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 516–520.

[8] Y.-J. Luo, K. Agres, and D. Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[9] Y.-N. Hung, I.-T. Chiang, Y.-A. Chen, and Y.-H. Yang, "Musical composition style transfer via disentangled timbre representations," in *Proc. of 28th International Joint Conference on Artificial Intelligence (IJCAI)*, Macao, China, 2020.

[10] G. Hadjeres, F. Nielsen, and F. Pachet, "GLSR-VAE: Geodesic latent space regularization for variational autoencoder architectures," in *Proc. of IEEE Symposium Series on Computational Intelligence (SSCI)*, Hawaii, USA, 2017, pp. 1–7.

[11] A. Pati and A. Lerch, "Latent space regularization for explicit control of musical attributes," in *Proc. of ICML Workshop on Machine Learning for Music Discovery Workshop (ML4MD), Extended Abstract*, Long Beach, California, USA, 2019.

[12] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam, "Disentangled multidimensional metric learning for music similarity," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6–10.

[13] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in *Proc. of 5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[14] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating Sources of Disentanglement in Variational Autoencoders," in *Advances in Neural Information Processing Systems 31 (NeurIPS)*, Montréal, Canada, 2018.

[15] H. Kim and A. Mnih, "Disentangling by Factorising," in *Proc. of 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweeden, 2018.

[16] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational Inference of Disentangled Latent Concepts from Unlabeled Observations," in *Proc. of 5th International Conference of Learning Representations (ICLR)*, Toulon, France, 2017.

[17] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems 27 (NeurIPS)*, Montréal, Canada, 2014.

[18] N. Siddharth, B. Paige, J.-W. van de Meent, A. Desmaison, N. D. Goodman, P. Kohli, F. Wood, and P. H. Torr, "Learning disentangled representations with semi-supervised deep generative models," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, Long Beach, California, USA, 2017.

[19] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem, "Disentangling factors of variations using few labels," in *Proc. of 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

[20] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader Networks:Manipulating Images by Sliding Attributes," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, Long Beach, California, USA, 2017, pp. 5967–5976.

[21] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep Convolutional Inverse Graphics Network," in *Advances in Neural Information Processing Systems 28 (NeurIPS)*, Montréal, Canada, 2015, pp. 2539–2547.

[22] C. Donahue, Z. C. Lipton, A. Balsubramani, and J. McAuley, "Semantically Decomposing the Latent Spaces of Generative Adversarial Networks," in *Proc. of 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[23] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dSprites: Disentanglement testing Sprites dataset," https://github.com/deepmind/dsprites-dataset, 2017, last accessed, 2nd April 2020.

[24] C. Burgess and K. Hyunjik, "3d-shapes Dataset," https://github.com/deepmind/3d-shapes, Feb. 2020, last accessed, 2nd April 2020.

[25] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3D Chairs: Exemplar Part-based 2D-3D Alignment using a Large Dataset of CAD Models," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio, USA, 2014, pp. 3762–3769.

[26] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee, "Deep Visual Analogy-Making," in *Advances in Neural Information Processing Systems 28 (NeurIPS)*, Montréal, Canada, 2015, pp. 1252–1260.

[27] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations," in *Proc. of 36th International Conference on Machine Learning (ICML)*, Long Beach, California, USA, 2019.

[28] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. of 2nd International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014.

[29] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," in *NIPS Workshop on Learning Disentangled Representations*, Long Beach, California, USA, 2017.

[30] P. Rubenstein, B. Scholkopf, and I. Tolstikhin, "Learning Disentangled Representations with Wasserstein Auto-Encoders," in *Proc. of 6th International Conference on Learning Representations (ICLR), Workshop Track*, Vancouver, Canada, 2018.

[31] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 29 (NeurIPS)*, Barcelona, Spain, 2016, pp. 2172–2180.

[32] M. Connor and C. Rozell, "Representing closed transformation paths in encoded network latent space," in *Proc. of 34th AAAI Conference on Artificial Intelligence*, New York, USA, 2020.

[33] J. Engel, M. Hoffman, and A. Roberts, "Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models," in *Proc. of 5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[34] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-Level Variational Autoencoder: Learning Disentangled Representations From Grouped Observations," in *Proc. of 32nd AAAI Conference on Artificial Intelligence*, New Orleans, USA, 2018.

[35] M. W. Gondal, M. Wüthrich, Đ. Miladinović, F. Locatello, M. Breidt, V. Volchkov, J. Akpo, O. Bachem, B. Schölkopf, and S. Bauer, "On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019, pp. 15 740–15 751.

[36] M. Bretan and L. Heck, "Learning semantic similarity in music via self-supervision," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[37] S. Gururani, A. Lerch, and M. Bretan, "A comparison of music input domains for self-supervised feature learning," in *Proc. of ICML Workshop on Machine Learning for Music Discovery Workshop (ML4MD), Extended Abstract*, Long Beach, California, USA, 2019.

[38] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proc. of 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 141–149.

[39] S. Lattner, M. Dörfler, and A. Arzt, "Learning complex basis functions for invariant representations of audio," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[40] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 3730–3738.

[41] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," in *Proc. of 35th*

International Conference on Machine Learning (ICML), Stockholm, Sweeden, 2018.

[42] G. Toussaint, "A Mathematical Analysis of African, Brazilian and Cuban Clave Rhythms," in *Proc. of BRIDGES: Mathematical Connections in Art, Music and Science*, 2002, pp. 157–168.

[43] M. S. Cuthbert and C. Ariza, "music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data," in *Proc. of 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, 2010.

[44] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: A steerable model for Bach chorales generation," in *Proc. of 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1362–1371.

[45] A. Pati, A. Lerch, and G. Hadjeres, "Learning to Traverse Latent Spaces for Musical Score Inpainting," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[46] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved Variational Inference with Inverse Autoregressive Flow," in *Advances in Neural Information Processing Systems 29 (NeurIPS)*, Barcelona, Spain, 2016, pp. 4743–4751.

[47] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of 3rd International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015.

[48] K. Ridgeway and M. C. Mozer, "Learning Deep Disentangled Embeddings With the F-Statistic Loss," in *Advances in Neural Information Processing Systems 31 (NeurIPS)*, Montréal, Canada, 2018, pp. 185–194.

[49] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, Long Beach, California, USA, 2017, pp. 6306–6315.

[50] T. Akama, "Controlling Symbolic Music Generation Based On Concept Learning From Domain Knowledge," in *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

DEEP MUSIC ANALOGY VIA LATENT REPRESENTATION DISENTANGLEMENT

Ruihan Yang[1] **Dingsu Wang**[1] **Ziyu Wang**[1] **Tianyao Chen**[1] **Junyan Jiang**[1,2] **Gus Xia**[1]

[1] Music X Lab, NYU Shanghai [2] Machine Learning Department, Carnegie Mellon University

[1]{ry649,dw1920,zz2417,tc2709,gxia}@nyu.edu, [2]junyanj@cs.cmu.edu

ABSTRACT

Analogy-making is a key method for computer algorithms to generate both natural and creative music pieces. In general, an analogy is made by partially transferring the music abstractions, i.e., high-level representations and their relationships, from one piece to another; however, this procedure requires *disentangling* music representations, which usually takes little effort for musicians but is non-trivial for computers. Three sub-problems arise: extracting latent representations from the observation, disentangling the representations so that each part has a unique semantic interpretation, and mapping the latent representations back to actual music. In this paper, we contribute an explicitly-constrained conditional variational autoencoder (EC^2-VAE) as a unified solution to all three sub-problems. We focus on disentangling the *pitch* and *rhythm* representations of 8-beat music clips conditioned on chords. In producing music analogies, this model helps us to realize the imaginary situation of "*what if*" a piece is composed using a different pitch contour, rhythm pattern, or chord progression by borrowing the representations from other pieces. Finally, we validate the proposed disentanglement method using objective measurements and evaluate the analogy examples by a subjective study.

1 Introduction

For intelligent systems, an effective way to generate high-quality art is to produce analogous versions of existing examples [15]. In general, two systems are analogous if they share common abstractions, i.e., high-level representations and their relationships, which can be revealed by the paired tuples *A : B :: C : D* (often spoken as A is to B as C is to D). For example, the analogy "the hydrogen atom is like our solar system" can be formatted as *Nucleus : Hydrogen atom :: Sun : Solar system*, in which the shared abstraction is "a bigger part is the center of the whole system." For generative algorithms, a clever shortcut is to make analogies by solving the problem of "*A : B :: C : ?*". In the context of music generation, if A is the rhythm pattern of a very lyrical piece B, this analogy can help us realize the imaginary situation of "what if B is composed with a rather rapid and syncopated rhythm C" by preserving the pitch contours and the intrinsic relationship between pitch and rhythm. In the same fashion, other types of "what if" compositions can be created by simply substituting A and C with different aspects of music (e.g., chords, melody, etc.).

A great advantage of *generation via analogy* is the ability to produce both *natural* and *creative* results. Naturalness is achieved by reusing the representations (high-level concepts such as "image style" and "music pitch contour") of human-made examples and the intrinsic relationship between the concepts, while creativity is achieved by recombining the representations in a novel way. However, making meaningful analogies also requires *disentangling* the representations, which is effortless for humans but non-trivial for computers. We already see that making analogies is essentially transferring the abstractions, not the observations — simply copying the notes or samples from one piece to another would only produce a casual re-mix, not an analogous composition [11].

In this paper, we contribute an explicitly-constrained conditional variational autoencoder (EC^2-VAE), a conditional VAE with explicit semantic constraints on intermediate outputs of the network, as an effective tool for learning disentanglement. To be specific, the encoder extracts latent representations from the observations; the semantic constraints disentangle the representations so that each part has a unique interpretation, and the decoder maps the disentangled representations back to actual music while preserving the intrinsic relationship between the representations. In producing analogies, we focus on disentangling and transferring the *pitch* and *rhythm* representations of 8-beat music clips when chords are given as the condition (an extra input) of the model. We show that EC^2-VAE has three desired properties as a generative model. First, the disentanglement is *explicitly coded*, i.e., we can specify which latent dimensions denote which semantic factors in the model structure. Second, the disentanglement does not sacrifice much of the reconstruction. Third, the learning does not require any analogous examples in the training phase, but the model is capable of making analogies in the inference phase. For evaluation, we propose a new metric and conduct a survey. Both objective and subjective evaluations show that our model significantly outperforms the baselines.

2 Related Work

2.1 Generation Via Analogy

The history of generation via analogy can trace back to the studies of non-parametric "image analogies" [15] and "playing Mozart by analogy" using case-based reasoning [29]. With recent breakthroughs in artificial neural networks, we see a leap in the quality of produced analogous examples using deep generative models, including music and image style transfer [7, 13], image-to-image translation [18], attribute arithmetic [3], and voice impersonation [12].

Here, we distinguish between two types of analogy algorithms. In a *broad* sense, an analogy algorithm is any computational method capable of producing analogous versions of existing examples. A common and relatively easy approach is supervised learning, i.e., to directly learn the mapping between analogous items from labeled examples [18, 27]. This approach requires little representation learning but needs a lot of labeling effort. Moreover, supervised analogy does not generalize well. For example, if the training analogous examples are all between lyrical melodies (the source domain) and syncopated melodies (the target domain), it would be difficult to create other rhythmic patterns, much less the manipulation of pitch contours. (Though improvements [1, 21, 32] have been made, weak supervision is still needed to specify the source and target domains.) On the other hand, a *strict* analogy algorithm requires not only learning the representations but also disentangling them, which would allow the model to make domain-free analogies via the manipulation of any disentangled representations. Our approach belongs to this type.

2.2 Representation Learning and Disentanglement

Variational auto-encoders (VAEs) [22] and generative adversarial networks (GANs) [14] are so far the two most popular frameworks for music representation learning. Both use encoders (or discriminators) and decoders (or generators) to build a bi-directional mapping between the distributions of observation x and latent representation z, and both generate new data via sampling from $p(z)$. For music representations, VAEs [2,9,24,30] have been a more successful tool so far compared with GANs [31], and our model is based on the previous study [30].

The motivation of representation disentanglement is to better interpret the latent space generated by VAE, connecting certain parts of z to semantic factors (e.g., age for face images, or rhythm for melody), which would enable a more controllable and interactive generation process. InfoGAN [5] disentangles z by encouraging the mutual information between x and a subset of z. β-VAE [16] and its follow-up studies [4,20,30] imposed various extra constraints and properties on $p(z)$. However, the disentanglement above are still *implicit*, i.e., though the model separates the latent space into subparts, we cannot define their meanings beforehand and have to "check it out" via *latent space traversal* [3]. In contrast, the disentanglement in Style-based GAN [19], Disentangled Sequential Autoencoder [23], and our EC^2-VAE are *explicit*, i.e., the meanings of different parts of z are defined by the model structure, so that the controlled generation is more precise and straightforward. The study Disentangled Sequential Autoencoder [23] is most related to our work and also deals with sequential inputs. Using a partially time-invariant encoder, it can approximately disentangle dynamic and static representations. Our model does not directly constrain z but applies a loss to intermediate outputs associated with latent factors. Such an indirect but explicit constraint enables the model to further disentangle the representation into pitch, rhythm, and any semantic factors whose observation loss can be defined. As far as we know, this is the first disentanglement learning method tailored for music composition.

3 Methodology

In this section, we introduce the data representation and model design in detail. We focus on disentangling the latent representations of pitch and rhythm, the two fundamental aspects of composition, over the duration of 8-beat melodies. All data come from the Nottingham dataset [10], regarding a $\frac{1}{4}$ beat as the shortest unit.

3.1 Data Representation

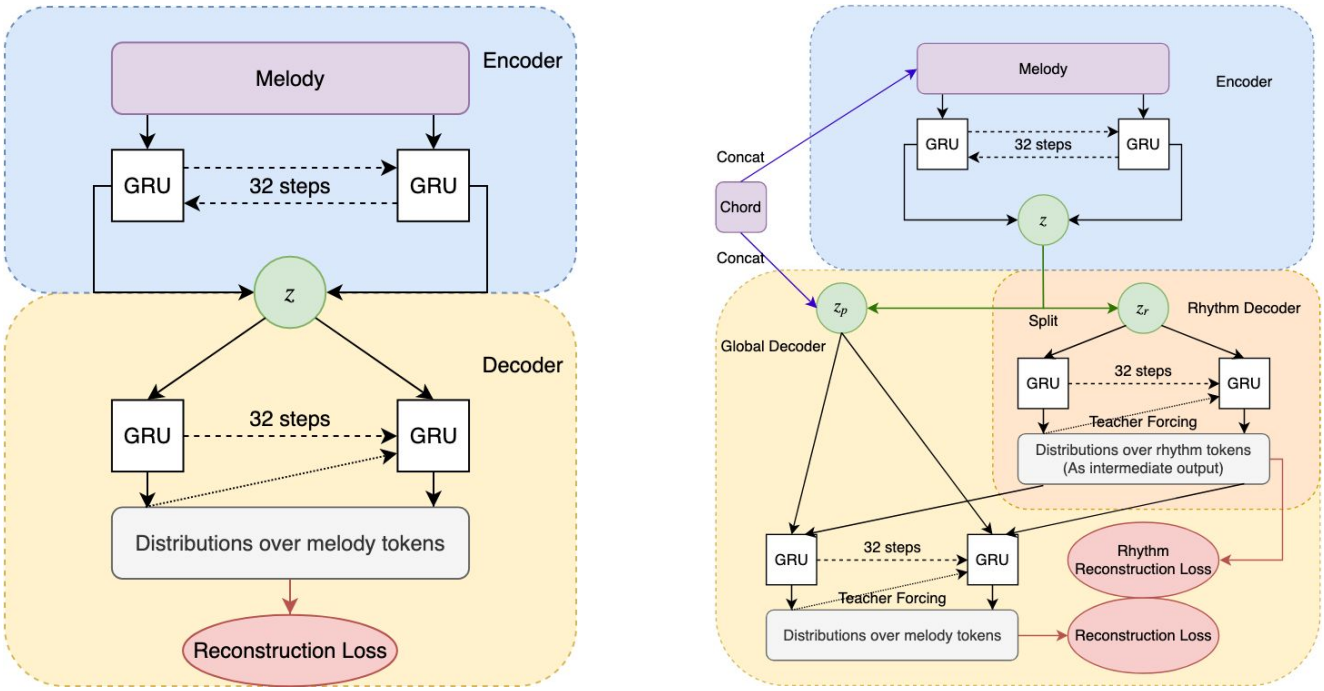
Each 8-beat melody is represented as a sequence of 32 one-hot vectors each with 130 dimensions, where each vector denotes a $\frac{1}{4}$-beat unit. As in [24], the first 128 dimensions denote the *onsets* of MIDI pitches ranging from 0 to 127 with one unit of duration. The 129^{th} dimension is the *holding* state for longer note duration, and the last dimension denotes *rest*. We also designed a rhythm feature to constrain the intermediate output of the network. Each 8-beat rhythm pattern is also represented as a sequence of 32 one-hot vectors. Each vector has 3 dimensions, denoting: an onset of any pitch, a holding state, and rest.

Besides, chords are given as a condition, i.e., an extra input, of the model. The chord condition of each 8-beat melody is represented as a chromagram with equal length, i.e., 32 multi-hot vectors each with 12 dimensions, where each dimension indicates whether a pitch class is activated.

3.2 Model Architecture

Our model design is based on the previous studies of [24, 30], both of which used VAEs to learn the representations of fixed-length melodies. Figure 1 shows a comparison between the model architectures, where Figure 1(a) shows the model designed in [30] and Figure 1(b) shows the model design in this study. We see that both use bi-directional GRUs [6] (or LSTMs [17]) as the encoders (in blue) to map each melody observation to a latent representation z, and both use uni-directional GRUs (or LSTMs) (with teacher forcing [26] in the training phrase) as the decoders (in yellow) to reconstruct melodies from z.

The key innovation of our model design is to assign a part of the decoder (in orange) with a specific subtask: to disentangle the latent rhythm representation z_r from the overall z by explicitly encouraging the intermediate output of z_r to match the rhythm feature of the melody. The other



(a) Vanilla sequence VAE.

(b) EC^2-VAE model.

Figure 1: A comparison between vanilla sequence VAE [30] and our model with condition and disentanglement.

part of z is therefore everything but rhythm and interpreted as the latent pitch representation, z_p. Note that this explicitly coded disentanglement technique is quite flexible — *we can use multiple subparts of the decoder to disentangle multiple semantically interpretable factors of z simultaneously* as long as the intermediate outputs of the corresponding latent factors can be defined, and the model shown in Figure 1(b) is the simplest case of this family.

It is also worth noting that the new model uses chords as a condition for both the encoder and decoder. The advantage of chord conditioning is to free z from storing chord-related information. In other words, the pitch information in z is "detrended" by the underlying chord for better encoding and reconstruction. The cost of this design is that we cannot learn a latent distribution of chord progressions.

3.2.1 *Encoder*

A single layer bi-directional GRU with 32 time steps is used to model $Q_\theta(z|x,c)$, where x is the melody input, c is the chord condition, and z is the latent representation. Chord conditions are given by concatenating with the input at each time step.

3.2.2 *Decoder*

The global decoder models $P_\phi(x|z,c)$ by multiple layers of GRUs, each with 32 steps. For disentanglement, z is splitted into two halves z_p and $z_r (z = \text{concat}[z_r, z_p])$, each being a 128-dimensional vector. As a subpart of the global decoder, the rhythm decoder models $P_{\phi_r}(r(x)|z)$ by a single layer GRU, where $r(x)$ is the rhythm feature of the melody. Meanwhile, the rhythm is concatenated with z_p and chord condition as the input of the rest of the global decoder to reconstruct the melody. We used cross-entropy loss for both rhythm and melody reconstruction. Note that the overall decoder is supposed to learn non-trivial relationships between pitch and rhythm, rather than naively cutting a pitch contour by a rhythm pattern.

3.3 Theoretical Justification of the ELBO Objective with Disentanglement

One concern about representation disentanglement techniques is that they sometimes sacrifice reconstruction power [20]. In this section, we prove that our model does not suffer much of the disentanglement-reconstruction paradox, and the likelihood bound of our model is close to that of the original conditional VAE, and in some cases, equal to it.

Recall the Evidence Lower Bound (ELBO) objective function used by a typical conditional VAE [8] constraint on input sample x with condition c:

$$\begin{aligned}\text{ELBO}(\phi,\theta) &= \mathbb{E}_Q[\log P_\phi(x|z,c)] \\ &\quad - \mathbb{KL}[Q_\theta(z|x,c)||P_\phi(z|c)] \le \log P_\phi(x|c)\end{aligned}$$

For simplicity, $\mathcal{D}$ denotes $\mathbb{KL}[Q_\theta(z|x,c)||P_\phi(z|c)]$ in the rest of this section. If we see the intermediate rhythm output in Figure 1(b) as hidden variables of the whole network, the new ELBO objective of our model only adds the rhythm reconstruction loss based on the original one, resulting in a lower bound of the original ELBO. Formally,

$$\begin{aligned}&\text{ELBO}^{\text{new}}(\phi,\theta) \\ &= \mathbb{E}_Q[\log P_\phi(x|z,c)] - \mathcal{D} + \mathbb{E}_Q[\log P_{\phi_r}(r(x)|z_r)] \\ &= \text{ELBO}(\phi,\theta) + \mathbb{E}_Q[\log P_{\phi_r}(r(x)|z_r)]\end{aligned}$$

where ϕ_r denotes parameters of the rhythm decoder. Clearly, ELBO^{new} is a lower bound of the original ELBO because $\mathbb{E}_Q[\log P_{\phi_r}(r(x)|z_r)] \le 0$.

Moreover, if the rest of global decoder takes the original rhythm rather than the intermediate output of rhythm decoder as the input, the objective can be rewritten as:

$$\begin{aligned}
&\text{ELBO}^{\text{new}}(\phi,\theta)\\
&=\mathbb{E}_Q[\underbrace{\log P_\phi(x|r(x),z_p,c)+\log P_\phi(r(x)|z_r,c)}_{\text{with } x \perp\!\!\!\perp z_r|r(x),c \text{ and } r(x) \perp\!\!\!\perp z_p|z_r,c}]-\mathcal{D}\\
&=\mathbb{E}_Q[\log P_\phi(x,r(x)|z,c)]-\mathcal{D}\\
&=\mathbb{E}_Q[\log P_\phi(x|z,c)+\log P_\phi(r(x)|x,z,c)]-\mathcal{D}\\
&=\text{ELBO}(\phi,\theta)
\end{aligned}$$

The second equal sign holds for a perfect disentanglement, and the last equal sign holds since $r(x)$ is decided by x, i.e., $P_\phi(r(x)|x,z,c)=1$. In other words, we show that under certain assumptions ELBO$^{\text{new}}$ with disentanglement is identical to the ELBO.

4 Experiments

We present the objective metrics to evaluate the disentanglement in Section 4.1, show several representative examples of generation via analogy in Section 4.2, and use subjective evaluations to rate the artistic aspects of the generated music in Section 4.3.

4.1 Objective Measurements

Upon a successful pitch-rhythm disentanglement, any changes in pitch of the original melody should not affect the latent rhythm representation much, and vice versa. Following this assumption, we developed two measurements to evaluate the disentanglement: 1) Δz after transposition, which is more qualitative, and 2) F-score of an augmentation-based query, which is more quantitative.

4.1.1 Visualizing Δz after transposition

We define F_i as the operation of transposing all the notes by i semitones, and use the L_1-norm to measure the change in z. Figure 2 shows a comparison between $\Sigma|\Delta z_p|$ and $\Sigma|\Delta z_r|$ when we apply F_i to a randomly chosen piece (where $i \in [1,12]$) while keeping the rhythm and underlying chord unchanged.

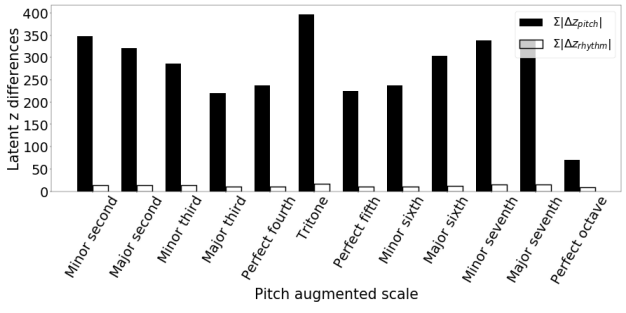


Figure 2: A comparison between Δz_p and Δz_r after transposition.

Here, the black bars stand for $\Sigma|\Delta z_p|$ and the white bars stand for the $\Sigma|\Delta z_r|$. It is conspicuous that when augmenting pitch, the change of z_p is much larger than the change of z_r, which well demonstrates the success of the disentanglement.

It is also worth noting that the change of z_p to a certain extent *reflects human pitch perception*. Given a chord, the change in z_p can be understood as the "burden" (or difficulty) to memorize (or encode) a transposed melody. We see that such burden is large for tritone (very dissonant), relatively small for major third, perfect fourth & fifth (consonant), and very small for perfect octave.

Due to the space limit, we only show the visualization of the latent space when changing the pitch. According to the data representation in Section 3.1, changing the rhythm feature of a melody would inevitably affect the pitch contour, which would lead to complex behavior of the latent space hard to interpret visually. We leave the discussion for future work but will pay more attention to the effect of the rhythm factor in Section 4.3.

4.1.2 F-score of Augmentation-based Query

The explicitly coded disentanglement enables a new evaluation method from an *information-retrieval* perspective. We regard the pitch-rhythm split in z defined by the model structure as the *reference* (the ground truth), the operation of factor-wise data augmentation (keeping the rhythm and only changing pitch randomly, or vice versa) as a *query* in the latent space, and the actual latent dimensions having the largest variance caused by augmentation as the *result set*. In this way, we can quantitatively evaluate our model in terms of precision, recall, and F-score.

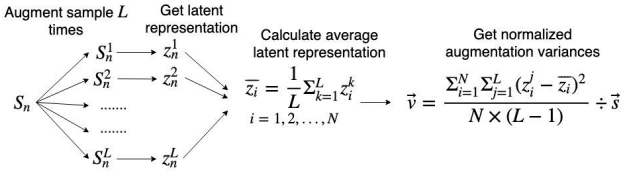


Figure 3: Evaluating the disentanglement by data augmentation.

	Pitch			Rhythm		
	pre.	rec.	F-s.	pre.	rec.	F-s.
EC^2-VAE	**0.88**	**0.88**	**0.88**	**0.80**	**0.80**	**0.80**
Random	0.5	0.5	0.5	0.5	0.5	0.5

Table 1: The evaluation results of pitch- and rhythm-wise augmentation-based query.

Figure 3 shows the detailed query procedure, which is a modification of the evaluation method in [20]. After pitch or rhythm augmentation for each sample, $\vec{v}$ is calculated as the average (across the samples) variance (across augmented versions) of the latent representations, normalized by the total sample variance $\vec{s}$. Then, we choose the first half (128 dimensions) with the largest variances as the result set. This precision, recall and F-score of this augmentation-based query result is shown in Table 1. (Here, precision and recall are identical since the size of the result set equals the dimensionality of z_p and z_r.) As this is the first tailored metric for explicitly coded disentanglement, we use random guess as our baseline.

4.2 Examples of Generation via Analogy

We present several representative "what if" examples by swapping or interpolating the latent representations of different pieces. Throughout this section, we use the following example (shown in Figure 4), an 8-beat melody from the Nottingham Dataset [10] as the source, and the target rhythm or pitch will be borrowed from other pieces. (MIDI demos are available at `https://github.com/cdyrhjohn/Deep-Music-Analogy-Demos`.)

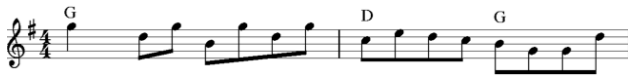
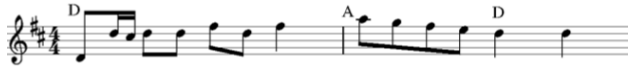


Figure 4: The source melody.

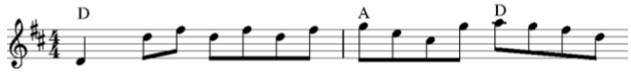
4.2.1 Analogy by replacing z_p

Two examples are presented. In both cases, the latent pitch representation and the chord condition of the source melody are replaced with new ones from other pieces. In other words, the model answers the analogy question: *"source's pitch : source melody :: target's pitch : ?"*

Figure 5 shows the first example, where Figure 5(a) shows the piece from which the pitch and chords are borrowed, and Figure 5(b) shows the generated melody. From Figure 5(a), we see the target melody is in a different key (D major) with a larger pitch range than the source and a big pitch jump in the beginning. From Figure 5(b), we see the generated new melody captures such pitch features while keeping the rhythm of the source unchanged.

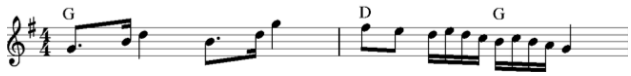


(a) Target's pitch and chord.

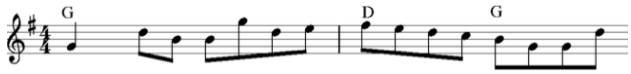


(b) The generated target music, using the pitch and chord from (a) and the rhythm from the source.

Figure 5: The 1st example of analogy via replacing z_p.



(a) Target's pitch and chord.



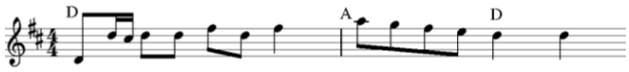
(b) The generated target, using the pitch and chord from (a) and the rhythm from the source.

Figure 6: The 2nd analogy example via replacing z_p.

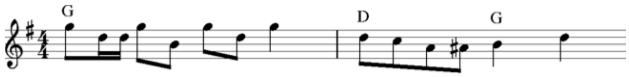
Figure 6 shows another example, whose subplots share the same meanings with the previous one. From Figure 6(a), we see the first measure of the target's melody is a broken chord of Gmaj, while the second measure is the G major scale. From Figure 6(b), we see the generated new melody captures these pitch features. Moreover, it retains the source's rhythm and ignores the dotted eighth and sixteenth notes in Figure 6(a).

4.2.2 Analogy by replacing z_r

Similar to the previous section, this section shows two example answers to the question: "*source's rhythm : source melody :: target's rhythm : ?*" by replacing z_r. Figure 7 shows the first example, where Figure 7(a) contains the new rhythm pattern quite different from the source, and Figure 7(b) is the generated target. We see that Figure 7(b) perfectly inherited the new rhythm pattern and made minor but novel modifications based on the source's pitch.

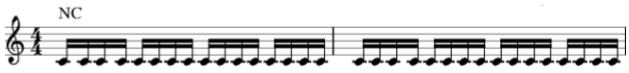


(a) Target's rhythm pattern.

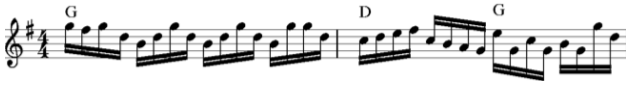


(b) The generated target music, using the rhythm of (a) while keeping source's pitch and chord.

Figure 7: The 1st example of analogy via replacing z_r.



(a) Target's rhythm pattern.

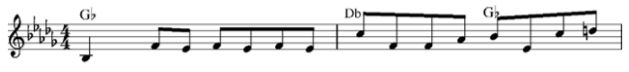


(b) The generated target music, using the rhythm of (a) while keeping source's pitch and chord.

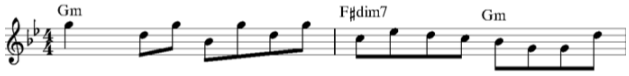
Figure 8: The 2nd analogy example via replacing z_r.

Figure 8 shows a more extreme case, in which Figure 8(a) contains only 16th notes of the same pitch. Again, we see the generated target in Figure 8(b) maintains the source's pitch contour while matching the given rhythm pattern.

4.2.3 Analogy by Replacing Chord



(a) Changing all the chords down a semitone, resulting in the key change from G major to Bb minor.



(b) Changing the key from G major to G minor.

Figure 9: Two examples of replacing the original chord.

Though chord is not our main focus, here we show two analogy examples in Figure 9 to answer "what if" the source melody is composed using some other chord progressions. Figure 9(a) shows an example where the key is Bb minor. An interesting observation is the new melody

contour indeed adds some reasonable modification (e.g. flipping the melody) rather than simply transposing down all the notes. It brings us a little sense of Jazz. Figure 9(b) shows an example where the key is changed from G major to G minor. We see melody also naturally transforms from major mode to minor mode.

4.2.4 *Two-way Pitch-Rhythm Interpolation*

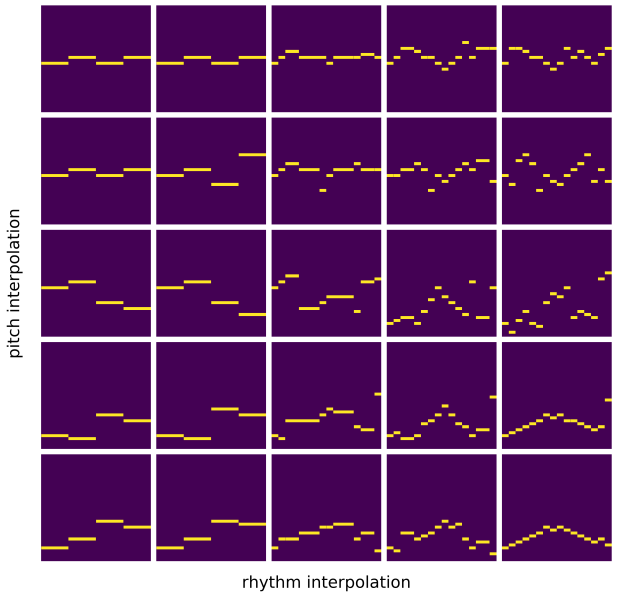


Figure 10: An illustration of two-way interpolation.

The disentanglement also enables a smooth transition from one music to another. Figure 10 shows an example of two-way interpolation, i.e., a traversal over a subspace of the learned latent representations z_r and z_p along 2 axes respectively, while keeping the chord as NC (no chord). Here, each square is a piano-roll of an 8-beat music. The top-left (source) and bottom-right (target) squares are two samples created manually and everything else is generated by interpolation using SLERP [28]. Note that the rhythmic changes are primarily observed moving along the "rhythm interpolation" axis, and likewise for pitch and the vertical "pitch interpolation" axis.

4.3 Subjective Evaluation

Besides objective measurement, we conducted a subjective survey to evaluate the quality of generation via analogy. We focus on changing the rhythm factors of existing music since this operation leads to an easier identification of the source melodies.

Each subject listened to two groups of five pieces each. All the pieces had the same length (64 beats at 120 bpm). Within each group, one piece was an original, human-composed piece from the Nottingham dataset, having a lyrical melody consisting of longer notes. The remaining four pieces were variations upon the original with more rapid rhythms consisting of 8$^{\text{th}}$ and 16$^{\text{th}}$ notes. Two of the variations were produced in a rule-based fashion by naively cutting the notes in the original into shorter subdivisions, serving as the *baseline*. The other two variations were generated with our EC^2-VAE by merging the z_p of the original piece and the z_r decoded from two pieces having the same rhythm pattern as the baselines but with all notes replaced with "do" (similar to Figure 8(a)). The subjects always listened to the original first, and the order of the variations was randomized. In sum, we compare three versions of music: 1) the original piece, 2) the variation created by the baseline, and, 3) the variation created by our algorithm. The subjects were asked to rate each sample on a 5-point scale from 1 (very low) to 5 (very high) according to three criteria:

1. *Creativity*: how creative the composition is.
2. *Naturalness*: how human-like the composition is.
3. *Overall musicality*.

A total of 30 subjects (16 female and 14 male) participated in the survey. Figure 11 shows the results, where the heights of bars represent means of the ratings the and error bars represent the MSEs computed via within-subject ANOVA [25]. The result shows that our model performs significantly better than the rule-based baseline in terms of creativity and musicality ($p < 0.05$), and marginally better in terms of naturalness. Our proposed method is even comparable to the original music in terms of creativity, but remains behind human composition in terms of the other two criteria.

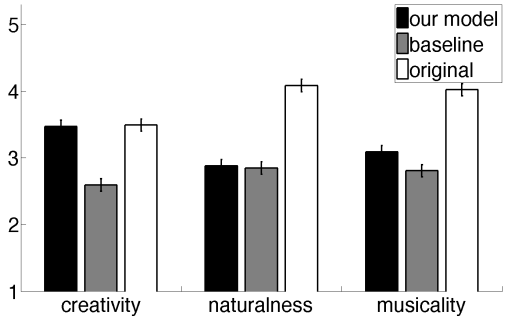


Figure 11: Subjective evaluation results.

5 Conclusion

In conclusion, we contributed an explicitly-constrained conditional variational autoencoder (EC^2-VAE) as an effective disentanglement learning model. This model generates new music via making analogies, i.e., to answer the imaginary situation of "what if" a piece is composed using different pitch contours, rhythm patterns, and chord progressions via replacing or interpolating the disentangled representations. Experimental results showed that the disentanglement is successful and the model is able to generate interesting and musical analogous versions of existing music. We see this study a significant step in music understanding and controlled music generation. The model also has the potential to be generalized to other domains, shedding light on the general scenario of generation via analogy.

6 Acknowledgement

We thank Yun Wang, Zijian Zhou and Roger Dannenberg for the in-depth discussion on music disentanglement and analogy. This work is partially supported by the Eastern Scholar Program of Shanghai.

7 References

[1] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] Gino Brunner, Andres Konrad, Yuyi Wang, and Roger Wattenhofer. Midi-vae: Modeling dynamics and instrumentation of music with applications to style transfer. *arXiv preprint arXiv:1809.07600*, 2018.

[3] Shan Carter and Michael Nielsen. Using artificial intelligence to augment human intelligence. *Distill*, 2(12):e9, 2017.

[4] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. NIPS, 2018.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[7] Shuqi Dai, Zheng Zhang, and Gus G Xia. Music style transfer: A position paper. *arXiv preprint arXiv:1803.06841*, 2018.

[8] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[9] Philippe Esling, Axel Chemla-Romeu-Santos, and Adrien Bitton. Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, pages 23–27, 2018.

[10] E. Foxley. Nottingham database, 2011.

[11] Y Gao. *Towards neural music style transfer*. PhD thesis, Master Thesis, New York University. https://github.com/821760408-sp/the . . . , 2017.

[12] Yang Gao, Rita Singh, and Bhiksha Raj. Voice impersonation using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2506–2510. IEEE, 2018.

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[15] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM, 2001.

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

[20] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

[21] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[23] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *arXiv preprint arXiv:1803.02991*, 2018.

[24] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. *arXiv preprint arXiv:1803.05428*, 2018.

[25] Henry Scheffe. *The analysis of variance*, volume 72. John Wiley & Sons, 1999.

[26] Nikzad Benny Toomarian and Jacob Barhen. Learning a trajectory using adjoint functions and teacher forcing. *Neural Networks*, 5(3):473–484, 1992.

[27] Christopher J Tralie. Cover song synthesis by analogy. *arXiv preprint arXiv:1806.06347*, 2018.

[28] Alan Watt and Mark Watt. Advanced animation and bendering techniques. 1992.

[29] Gerhard Widmer and Asmir Tobudic. Playing mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research*, 32(3):259–268, 2003.

[30] Ruihan Yang, Tianyao Chen, Yiyi Zhang, and Gus Xia. Inspecting and interacting with meaningful music representations using vae. *arXiv preprint arXiv:1904.08842*, 2019.

[31] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

EVALUATION OF LATENT SPACE DISENTANGLEMENT IN THE PRESENCE OF INTERDEPENDENT ATTRIBUTES

Karn N. Watcharasupat[1,2] **Alexander Lerch**[1]
[1]Center for Music Technology, Georgia Institute of Technology, Atlanta, GA, USA
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
karn001@e.ntu.edu.sg, alexander.lerch@gatech.edu

ABSTRACT

Controllable music generation with deep generative models has become increasingly reliant on disentanglement learning techniques. However, current disentanglement metrics, such as mutual information gap (MIG), are often inadequate and misleading when used for evaluating latent representations in the presence of interdependent semantic attributes often encountered in real-world music datasets. In this work, we propose a dependency-aware information metric as a drop-in replacement for MIG that accounts for the inherent relationship between semantic attributes.

1. INTRODUCTION

Disentanglement learning has been an influential field of studies for controllable music generation with variational autoencoders (VAEs). A number of previous studies have attempted supervised disentanglement learning techniques on several semantic attributes such as rhythm, pitch range [1], note density, contour [2], arousal [3], style [4], and genre [5] to varying degrees of success. However, learning to simultaneously manipulate multiple attributes, in particular, remains a difficult task to both achieve and objectively evaluate [6] due to the limitation of current metrics.

One major issue with popular disentanglement metrics [7], such as mutual information gap (MIG) [8], separate attribute predictability (SAP) [9], and modularity [10], is that they were designed for independent generative factors, rather than real-world semantic attributes. As semantic attributes related to music are often highly interdependent, these metrics do not provide an accurate reflection of the 'quality' of learnt latent representation regularized for multiple interdependent attributes. Information inherently shared between attributes is penalized in the same way as that due to undesired entanglement issues.

In this work, we propose a dependency-aware metric based on mutual information (MI) to act as a drop-in replacement for MIG. Preliminary experiments were carried out to demonstrate the benefits of the proposed metrics over MIG.

2. PROPOSED METRICS

Consider a set of attributes $\{a_i\}_{i=1}^{M}$ and a latent vector $\boldsymbol{z} \in \mathbb{R}^D$ with $M \leq D$. Without loss of generality, for $i \leq M$, we assume z_i is regularized for a_i. The remaining dimensions are unregularized. $\mathcal{H}(\cdot)$ denotes entropy while $\mathcal{I}(\cdot,\cdot)$ denotes mutual information.

MIG was proposed in [8] to measure the degree of disentanglement in a latent space. The idea behind MIG can be said to measure: *for each attribute, the normalized difference between the mutual information between the attribute and its most informative latent dimension, and that between the attribute the second-most informative latent dimension*. Mathematically, MIG is given by

$$\text{MIG}(a_i) = \left(\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j)\right) / \mathcal{H}(a_i), \qquad (1)$$

where $j = \arg\max_{k \neq i} \mathcal{I}(a_i, z_k)$. It is reasonable to assume $i = \arg\max_k \mathcal{I}(a_i, z_k)$ in a supervised setting; otherwise MIG takes negative values to indicate regularization failure. The normalization is given by $\mathcal{H}(a_i)$, which would be the maximum possible difference in MI between a latent dimension z_i coding perfectly for a_i, i.e., $\mathcal{I}(a_i, z_i) = \mathcal{H}(a_i)$ and the second-most containing no information about a_i, i.e., $\mathcal{I}(a_i, z_j) = 0$. As such, MIG is bounded above by one.

However, given the interdependence of semantic attributes, if $j \leq M$, the ideal value of the difference $\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j)$ is no longer $\mathcal{H}(a_i)$ since

$$\mathcal{I}(z_j, a_j) > 0 \wedge \mathcal{I}(a_i, a_j) > 0 \implies \mathcal{I}(a_i, z_j) > 0. \quad (2)$$

For regularized latent dimensions, we consider a pair of inherently entangled attributes (a_i, a_j), i.e., $\mathcal{I}(a_i, a_j) > 0$. Under the ideal case where z_i is fully informative [7] about a_i, i.e., $\mathcal{H}(a_i|z_i) = 0$, we have

$$\begin{aligned} &\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j) \\ &\quad = [\mathcal{H}(a_i) - \mathcal{H}(a_i|z_i)] - [\mathcal{H}(a_i) - \mathcal{H}(a_i|z_j)] \quad (3) \\ &\quad = \mathcal{H}(a_i|z_j) \qquad \because \mathcal{H}(a_i|z_i) = 0. \quad (4) \end{aligned}$$

Moreover, in the ideal case, z_j and a_j also have an invertible mapping between each other, this means that

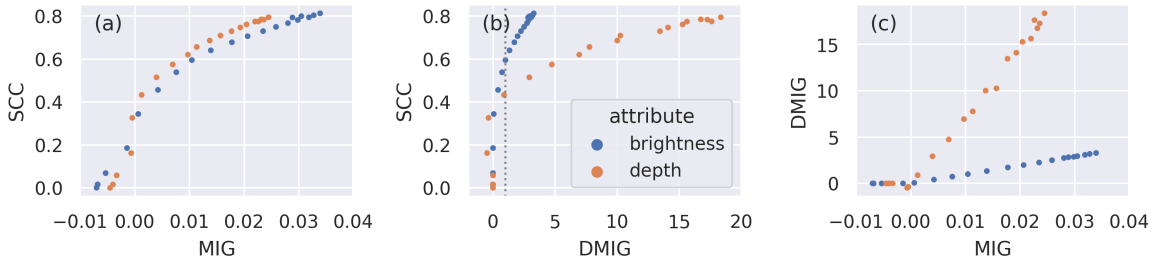


Figure 1: Plots of (a) SCC against MIG, (b) SCC against DMIG, and (c) DMIG against MIG, on the validation set.

$\mathcal{H}(a_i|z_j) = \mathcal{H}(a_i|a_j)$. Hence, in the ideal case, the difference is given by

$$\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j) = \mathcal{H}(a_i|a_j). \tag{5}$$

As such, we extend the definition of mutual information gap to the dependency-aware mutual information gap (DMIG) as follows

$$\text{DMIG}(a_i) = \begin{cases} \left(\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j)\right) / \mathcal{H}(a_i|a_j) & j \leq M \\ \left(\mathcal{I}(a_i, z_i) - \mathcal{I}(a_i, z_j)\right) / \mathcal{H}(a_i) & j > M, \end{cases} \tag{6}$$

where $j = \arg\max_{k \neq i} \mathcal{I}(a_i, z_k)$. DMIG remains faithful to the core idea of MIG but modifies the normalization to properly account for inter-attribute dependencies. When a_i and a_j are independent, $\mathcal{H}(a_i) - \mathcal{I}(a_i, a_j) = \mathcal{H}(a_i)$ and the DMIG reduces to vanilla MIG.

Note that in the case of continuous random variables, differential entropy can be negative, unlike discrete Shannon entropy. This is particularly evident with conditional differential entropy and may result in DMIG values above unity whenever $\mathcal{H}(a_i|z_i)/\mathcal{H}(a_i|z_j)$ is negative.

3. EXPERIMENTS

To illustrate the key features of the dependency-aware metrics, we evaluate the latent space of a VAE model trained to reconstruct raw musical audio while being regularized for two highly correlated attributes [1] .

3.1 Data and model

We use the NSynth dataset [11], which is a large-scale dataset of musical notes played by various instruments with diverse timbral qualities. The dataset provides 4-second snippets sampled at 16 kHz. From the raw audio provided by NSynth, we extract two semantic attributes, namely, *brightness* and *depth* using the AudioCommons Timbral Model [12]. Since both the brightness and depth features are heavily influenced by the spectral distribution of the sound [13], they are strongly correlated.

We trained a convolutional VAE model to reconstruct the log-magnitude spectrogram of the audio and obtain reconstructed time-domain audio using a phase-bypass reconstruction. The models are trained using the attribute-regularized β-VAE loss function [2, 14, 15]

$$\mathcal{L} = \mathcal{R}(\hat{\mathbf{x}}; \mathbf{x}) + \beta \mathcal{D}(\mathbf{z}) + \gamma \textstyle\sum_i \mathcal{A}(z_i; a_i), \tag{7}$$

where $\mathcal{R}(\cdot)$ is the reconstruction loss implemented via the mean square error on the log-magnitude spectrograms, $\mathcal{D}(\cdot)$ is the KL divergence term with a standard normal prior, and $\mathcal{A}(\cdot)$ is the AR-VAE regularization from [2]. We used $D = 512$, $\beta = 1$, and $\gamma = 10$.

3.2 Results

Figure 1 plots the MIG, DMIG, and Spearman correlation coefficient (SCC) of the attributes (brightness and depth) with respect to their respective regularized latent dimensions on the validation set over the course of the training. Due to the high correlation between brightness and depth, for most of the training, the most and second-most informative latent dimensions in MIG/DMIG are the regularized ones that encode for the attributes.

As seen from Figure 1(a), the MIG values are generally very low (in the order of 10^{-2}, out of maximum 1) despite the SCC indicating successful encoding of the attribute information into the latent dimension. This is due to the high mutual information between brightness and depth, resulting in a very low true bound for MIG. On the other hand, we can observe from Figure 1(b) that DMIG reflects more clearly the quality of the latent space as it encodes the attribute; the rapid improvement in SCC mostly occurred before DMIG reaches one (dotted line). In Figure 1(c), the highly linear relationship between MIG and DMIG further demonstrates the idea that DMIG is simply MIG renormalized to better reflect the dependencies between semantic attributes coded by the model. Admittedly, the peculiarities of differential conditional entropy and the practical computation of mutual information and entropy estimates [16] have contributed to a DMIG range that is much larger than vanilla MIG. We will be working to resolve this limitation in future work.

4. CONCLUSION

In this work, we propose a dependency-aware extension to a popular disentanglement metrics, mutual information gap (MIG), to better account for inter-attribute dependencies often observed in real-world datasets. Key features of the proposed dependency-aware MIG were demonstrated via an experiment on an audio dataset with highly correlated timbral attributes.

[1] See the supplementary materials for full experimental details at https://github.com/karnwatcharasupat/dependency-aware-mi-metrics.

5. ACKNOWLEDGEMENT

K. N. Watcharasupat acknowledges the support from the CN Yang Scholars Programme, Nanyang Technological University, Singapore.

6. REFERENCES

[1] A. Pati and A. Lerch, "Latent Space Regularization for Explicit Control of Musical Attributes," in *Extended Abstracts for ML4MD, ICML*, 2019.

[2] ——, "Attribute-based regularization of latent spaces for variational auto-encoders," *Neural Comput. Appl.*, vol. 33, no. 9, pp. 4429–4444, 2020.

[3] H. H. Tan and D. Herremans, "Music fadernets: Controllable music generation based on High-Level features via Low-Level feature modelling," in *Proc. ISMIR*, 2020.

[4] Y. N. Hung, I. T. Chiang, Y. A. Chen, and Y. H. Yang, "Musical composition style transfer via disentangled timbre representations," in *Proc. IJCAI*, 2019, pp. 4697–4703.

[5] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, "MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer," *Proc. ISMIR*, pp. 747–754, 2018.

[6] A. Pati and A. Lerch, "Is Disentanglement enough? On Latent Representations for Controllable Music Generation," in *Proc. ISMIR*, 2021.

[7] K. Do and T. Tran, "Theory and Evaluation Metrics for Learning Disentangled Representations," in *Proc. ICLR*, 2020.

[8] T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Proc. NeurIPS*, 2018.

[9] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," in *Proc. ICLR*, 2018.

[10] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the F-statistic loss," in *Proc. NeurIPS*, 2018, pp. 185–194.

[11] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proc. ICML*, 2017, pp. 1771–1780.

[12] A. Pearce, S. Safavi, T. Brookes, R. Mason, W. Wang, and M. Plumbley, "Deliverable D5.8 - Release of timbral characterisation tools for semantically annotating non-musical content," Audio Commons Initiative, Tech. Rep., 2019.

[13] ——, "Deliverable D5.2 - First prototype of timbral characterisation tools for semantically annotating non-musical content," Audio Commons Initiative, Tech. Rep., 2017.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. ICLR*, 2014.

[15] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2017.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.