The Commonwealth

Commonwealth Cyber Journal \ 75
ISSN 2959-3018 (print), ISSN 2959-3026 (online)
thecommonwealth.org/cyber-journal

Special Section on Artificial Intelligence

Crimes of Influence: Generative Artificial Intelligence-led Crime as a Service

Nicole Matejic[1] and Chris Wilson[2]

Abstract

'Crimes of Influence' – crimes that seek to influence people towards harmful outcomes – will be one of the defining features of generative artificial intelligence (AI)-led cybercrime. With an ability to persuade and influence at potentially unavoidable economies of scale, crimes of influence leverage the heuristics and biases that form part of everyday human cognition in ways that mislead, deceive, impair, disrupt, degrade and/or deny user-normative decision-making. Supported by evolving Crime as a Service (CaaS) models engineered to exploit human cognition, generative AI will challenge legislators, regulators and policymakers in ways that they are currently underprepared for. With generative AI able to surpass its initial deployment configuration via adaptive learning, as well as demonstrating unintended consequences, 'who' is then responsible for the crimes it commits when the only human touchpoints occur at the design, deployment and delivery of proceeds-of-crime stages? This paper draws upon emergent scholarship to explore the present-day exploration of generative AI tools by cybercriminals and terrorists, before looking to a hypothetical future and exploring successful initiatives attempting to address this challenge.

Keywords: cybercrime, generative artificial intelligence, crimes of influence, crime as a service (CaaS), terrorism, violence, behavioural economics, influence.

Introduction

The positive benefits of generative artificial intelligence (AI) are undeniable, particularly for science, commerce, education and healthcare innovation. However, like all forms of new and emergent technology, there will always be those who seek to harness its potential for malign and criminal purposes. This paper will contend that generative AI will

1 School of Terrorism and Security Studies, Charles Sturt University (New South Wales, Australia). Email: nmatejic@csu.edu.au
2 Faculty of Arts, Politics and International Relations, University of Auckland (Auckland, New Zealand).

enable criminals and other actors with malign intent to use Crime as a Service (CaaS) products engineered to exploit cognitive biases and heuristics in ways that mislead, deceive, impair, disrupt, degrade and/or deny user decision-making. This cognitive frontier pits coercive and deceptive generative AI against citizens going about their online and virtual lives with little to no understanding of how they may be influenced or nudged towards scammers, fraudsters and violent extremists, and potentially misled into conflicts and violence. We refer to this method as 'crimes of influence'.[3] In delivering new tools at significant economies of scale, generative AI will challenge society in ways we are currently underprepared for. This is particularly true of the relationship between crime and terrorism in transnational organised settings where hybridised groups with both political and criminal ambitions remain deeply intertwined (Makarenko, 2004). While the interplay between technology, society, crime and terrorism is not a new observation, generative AI poses new questions for governments, the technology sector and civil society, particularly 'who' is potentially responsible when generative AI is used as a conduit for cybercrime. This paper will consider how present-day cybercrime is evolving to exploit the opportunities generative AI presents, and how existing transnational organised crime and terrorist organisations are adapting in parallel. The paper begins with contextual definitions before exploring the role of influence in creating permissive environments for criminals and violent extremists. The following section explores the current landscape, followed by a look at future-state crimes of influence. The remainder of the paper considers the current state of governmental responses and multistakeholder thinking against a backdrop of rapid technological advancement.

Generative AI

Public awareness of, and interest in, AI has increased significantly following the launch of Open AI's public ChatGPT chatbot in November 2023. While AI is not a new technology, ChatGPT offers a first look at a seemingly 'magical' tool (Byrne, 2023; IBM, 2023) that is likely to change the way people search, find, consume and produce information. While AI chatbots like ChatGPT, Google's Gemini (formerly Bard) and Anthropic's Claude all rely on 'purpose-built (large language) models deployed for dedicated tasks' such as telling jokes and writing human-like essays (IBM, 2023), generative AI refers to a 'category of AI

3 The legal discipline considers a wide range of influence vectors and their effects on resulting crimes. For example: 'crimes of passion' considers whether a crime was committed in response to a provocation versus a premeditated act (Cornell Law School, 2023); the abuse of power and influence for personal gain is often a feature in bribery and corruption crimes (Keeler, 2024); the influence a serious mental disorder has on a person's culpability when they committed a crime is considered (Hartvigsson, 2023); and 'undue influence' in the context of contract and criminal law considers 'a situation in which one party has exerted pressure or influence over the other party, resulting in the weaker party being induced to enter into a contract' or action that is not in their best interest (UOLLB, 2024). In this paper we use the term 'crimes of influence' more broadly, although we acknowledge culpability regarding the underlying legal aspect of 'who' committed the crime could reasonably be challenged when a generative AI CaaS has evolved beyond its initial programming.

algorithms that generates new' content such as 'images, text, audio' (World Economic Forum, 2023), animations, 3D models and many other types of data (Nvidia, 2023).

When asked 'Tell me how generative AI will change cybercrime?' ChatGPT 3.5 helpfully defined AI before listing eight impacts: (1) advanced malware and phishing attacks (2) automated vulnerability exploitation (3) evasion of security measures (4) data forgery and deepfakes (5) enhanced social engineering (6) automated password cracking (7) zero-day exploits and (8) faster and more targeted (cyber) attacks. ChatGPT concluded by noting 'generative AI is not solely a tool for cybercriminals. It can also be used for cybersecurity purposes' (ChatGPT, 2023). Google's Gemini responded to the same query with a definition of AI before listing many of the same impacts as ChatGPT while adding money laundering, traditional financial and cryptocurrency frauds, and blackmail before concluding with a helpful three-point summary of how generative AI is being used to improve cybersecurity (Google Gemini, 2023). Anthrophic's Claude responded by acknowledging its limited perspective on the query given it 'is an emerging and complex issue' before listing many of the same impacts as ChatGPT and Gemini (Anthropic, 2023). What all the chatbots sampled indicated, however, was that the cybercrimes of the future are predicated on the technology's ability to influence, deceive and coerce susceptible people. While malign influence is not a new phenomenon, as this paper will explore in future sections, generative AI represents a new frontier for those deploying influence-based techniques that seek to generate specific, harmful effects.

Generative AI CaaS: Influence for sale at scale

CaaS enables individuals and/or organised criminal groups to buy the 'tools, infrastructure, and services' they need to commit their crimes for a fee. No technical skills are required on the part of the CaaS buyer with the vendor providing readymade services tailored to a marketplace that centres around common crime types such as malware-as-a-service, exploit-as-a-service (also known as zero-day exploits) and infrastructure-as-a-service (HKCERT, 2023).[4] CaaS is a multi-billion-dollar-a-year industry. Cybersecurity intelligence company Heimdal Security predicts that, over the next five years, the economic losses associated with cybercrime will increase 'by 23% per year, reaching a total of USD 23.84 trillion annually by 2027' (Chebac, 2023). These estimates do not specifically account for any generative AI-led activities within the cybercrime ecosystem. With such huge revenues at stake, it is perhaps unsurprising that CaaS has evolved to provide enterprise-grade 'product development, technical support, distribution, quality assurance and help desk' wrap-around

4 Malware-as-a-service that delivers malicious software, ransomware, spyware and trojans to a buyer seeking to infect targeted devices to steal sensitive information or hold data hostage. Exploit-as-a-service is a more niche offering, providing access to profitable, yet unknown, security vulnerabilities. Exploits can target a range of organisations, using their ability to exploit their cyber environments to steal data, money and information, to conduct espionage, and more. Infrastructure-as-a-service delivers networked solutions, such as botnets, which can be put to use to spam targets, host malicious or illegal content, or conduct denial-of-service attacks (HKCERT, 2023).

services' (Lewis, 2018). Few have considered how CaaS models will innovate alongside generative AI. However, it is not difficult to foresee how transnational crime organisations and terrorists (or nation states) could weaponise CaaS generative AI products to persuade and exploit. This is particularly true of social engineering-based cybercrime, which is predicated on deceiving and co-opting people into becoming unwitting victims.

Why generative AI will be the most persuasive human invention yet

Contemporary influence is both technological and psychosocial in nature. AI and generative AI enables influence activities of all kinds to be conducted at a low-cost, high-yield scale while the psychosocial element relies on priming and framing information in ways that engage particular cognitive processes (such as emotion, biases and heuristics) to influence decision-making. Extensive literature from the behavioural sciences, psychology, behavioural economics and neuropsychology disciplines broadly considers how influence occurs (Cialdini, 2016 Cialdini, 2021; Thaler and Sunstein, 2009; Kahneman, 2011; Wanless, 2017; Ariely, 2008; Nickerson, 1998).

Behavioural economists consider influence as often (but not always) the product of a person's decision-making environment. Thaler and Sunstein (2009) contend that choice architecture – the way decisions are presented – are constructed to influence people's choices towards defined, predictable outcomes. Influence, they argue, occurs due to a person's susceptibility to how varying biases and heuristics are engaged (Thaler and Sunstein, 2009). Cialdini (2021) refers to these environments as 'fixed-action patterns' involving 'intricate sequences of behaviour' noting that human automaticity is a well-known principle of human behaviour that is triggered by certain stimuli.

Psychologists refer to the way the brain approaches decision-making as System 1 and System 2 thinking. 'System 1 operates automatically and quickly with little or no effort and no sense of voluntary control' while System 2 'allocates attention to effortful mental activities... The operations of System 2 are often associated with the experiences of agency, choice and concentration' (Kahneman, 2011).[5] Inducing System 1 thinking, such as is the case in scams for example, in which emails or messages look authentic enough that the brain processes the content with little if any friction, is incredibly useful to cybercriminals. In fact, if they induce System 2 thinking, which provokes a level of consideration that includes attention to detail, it is more likely that their scam will be

5 See Kahneman, D. (2011) *Thinking Fast and Thinking Slow*, which goes into detail about System 1 and System 2 thinking. For illustrative purposes, thinking that can be attributed to System 1 includes 'detecting one object is closer than another, detecting hostility in a voice, understanding simple sentences and orienting to the source of a sudden sound'. System 2 thinking is more deliberate, such as 'focusing on the voice of a particular person in a crowded and noisy room, searching memory to identify a surprising sound, telling someone your phone number, filling out a tax form, and comparing two washing machines for overall value.'

detected. This approach is predicated on hijacking heuristics – the brain's way of making mental shortcuts in decision-making – allowing for faster cognition on what Kahneman (2011) refers to as 'simple procedures that help find adequate, though often imperfect, answers to difficult questions'. Cybercriminals may also seek to hijack cognitive biases, particularly in social-engineering settings where they nudge into systemic errors in decision-making. When combined with target motivation, 'non-random errors in thinking' result in 'judgements that deviate from what would be considered desirable' against benchmarks of social norms and logic (Ariely, 2008; Nickerson, 1998). While biases and heuristics often work in tandem, this is not always the case.

Understanding the mechanics of influence is particularly pertinent to online environments, which are participatory by design (Wanless, 2017) and are often incentivised towards a choice architect's desired outcomes. Algorithmic influence, for example, contributes to the choice architecture found on social media networks. While algorithms have come under increasing scrutiny for their ability to nudge people into ecosystems that contribute to polarisation and radicalisation, there are often 'competing logics' in play. While users may value the social aspects of online community, the network's priorities differ. From 'platform growth and revenue… extending use times and attracting advertisers' (Munn, 2020) to how algorithms balance this dichotomy is of continued interest to regulators and civil society, particularly as generative AI tools are recognised as both part of the problem and near-future solution to some of the most pressing content-moderation challenges of our time (Wolbers, 2023).

To further complicate how influence occurs, the brain's neurochemistry, specifically hormones and neuropeptides, also play a part. As Zak (2017) explains, oxytocin, the hormone that builds trust between a person and a stranger, 'is evolutionary old. This means that the trust and sociality that oxytocin enables are deeply embedded into our nature' (Zak, 2017). Dopamine and cortisol have also been observed to influence human decision-making. Simi et al. (2017) noted that dopamine's impact on cognition impairs a person's ability to think critically because it activates neural pathways in the brain's reward centre, similar to the way illicit drugs provide a high and become addictive. While Harms (2017) notes that cortisol responses can be induced in people which may lead to a decrease in cognitive flexibility. The effects of this, Harms (2017) explains, is that people rely too heavily on historical information or information they have only recently been exposed to, affecting their critical and future-oriented cognition. But influence is more than heuristics, biases and how information is pushed at people. Cialdini's (2016) concept of 'pre-suasion' considers how information can be front-loaded, ahead of time, enabling a person to draw together 'seemingly insignificant cues and unimportant details' which then primes them for influence at a later date (Cialdini, 2016).

Combined, these technological, psychosocial and neurochemical attributes create a permissive environment for susceptible minds. While not everyone will be susceptible to the triggers that cybercriminals or violent extremists often present, such as phishing, social-engineering exploits and radicalising content, the way that generative

AI produces increasingly visual content leverages the way that humans preference sight over all other senses (Enoch et al., 2019), even though it is unreliable (Synnott, 2022). When it comes to decision-making, generative AI has the potential to further exacerbate the perils found within System 1 thinking due to its ability to deliver realistic information visually. The creation of false images, audio and video, such as AI-generated deepfakes, requires little expertise and few resources making this widely accessible to a range of provocateurs. The technology needed for these tasks are often open source and available publicly, enabling the quick production of misleading or deceptive content. Further, visual content such as photographs, video and memes generate a great deal more engagement than simple text. It elicits greater emotion, is disseminated further and, therefore, has a greater impact on influencing people towards outcomes such as radicalisation towards extremism. It is at this cognitive juncture that generative AI will increase the effectiveness, and potentially pervasiveness, of crimes of influence.

With technology moving towards more immersive, augmented and virtual environments, 'biometric psychographics' perhaps presents the biggest opportunity for cybercriminals and terrorists to fully exploit human cognition. 'Biometric psychographics', a term coined by Heller (2021) to capture the convergence of 'traditional biometrics and predictive behavioural analysis', is increasingly becoming a part of everyday AI-based technology. With features such as eye tracking and pupil response, facial and vocal scanning, measuring galvanic skin response, EEG (brainwaves), ECG (pulse and blood pressure) as well as EMG (muscle tension), gait, facial expressions and more, everything from wearable technology to gaming, and augmented and virtual environments, is predicated on extracting biometric psychographics to, ostensibly, deliver better user experiences (Heller, 2021; Meta, 2021). The intimate nature of such data collection should give policy and lawmakers pause. While the legitimate use of the data is clear, the lack of neutrality in online environments raises concerns about privacy, security, algorithmic misuse and manipulation. Scholars have already cautioned that immersive technology (such as virtual reality) creates cognitive challenges that result in the brain processing and remembering experiences as if they had occurred in the real word – awakening spatial memory that quite literally 'draws on the brain in permanent ink' (Heller, 2021). For victims of generative AI-led cybercrime and terrorism, this could have catastrophic effects on their mental and physical health.

That society is not yet fully immersed in augmented and virtual worlds in which generative AI's capabilities can fully exploit biometric psychographics, provides an opportunity for these risks to be better understood and mitigated. The capacity to influence others towards harmful outcomes presents risks. These are explored in the following sections.

Accountability limitations

With criminal law resting on a burden of proof based on *actus reus*, the physical act of the crime, as well as *mens rea*, the mental intent to commit the crime, generative AI has the potential to deliver both intent and means without human intervention in ways

similar to today's smart legal contracts.[6] This is because generative AI has the capability to learn from its environment and to evolve autonomously beyond its initial directions. This could make CaaS enterprises particularly troublesome to prosecute. Whether by self-evolutionary adaptation or the product of an unexpected outcome, whether CaaS providers can ever truly retain complete control over their systems requires careful consideration. For example, well-engineered CaaS enterprises may leave the deployment and delivery-of-proceeds-of-crime stages as the only human touchpoints in the process. Whether or not those who ostensibly developed, procured, deployed and/or profited from the generative AI-led CaaS could be prosecuted as a party to such an activity remains to be seen. This is particularly so in an increasingly decentralised marketplace where the use of cryptocurrency and other forms of decentralisation that obfuscate identity are commonplace. It is plausible that, alongside self-adaptive exploit-driven innovation, generative AI CaaS models may also adapt to circumnavigate anti-money laundering frameworks, further driving the proceeds of crime into decentralised financial environments. Law and policymakers will also need to consider how they treat and manage persuasive and deceptive conduct by generative AI as a standalone entity. For example, persuasive generative AI 'could become better at predicting and nudging behaviour – becoming more capable of manipulation' and deception, including against those who programmed and deployed it (Hendrycks et al., 2023).

Disruption of CaaS marketplaces may be possible, such as via interventions that result in raising the cost of conducting generative AI-led CaaS business. In an environment where those systems are in a state of perpetual adaptation, however, such interventions are unlikely to fully disrupt or degrade their capabilities for long. It is possible they will have little to not effect at all, instead resulting in raising the cost and complexity of interventions by law enforcement and regulators. This leaves the economics of scale of generative AI-based CaaS models still weighed heavily in favour of cybercriminals and creates pre-hardened CaaS-permissive environments. Early learning from adaptive AI-based malware and precision-based social engineering (HYAS, 2023) in this regard could help researchers understand these challenges more fully.

Current landscape

The potential for crimes of influence to have negative effects on users can be seen in many of the cybercrime types identified by the aforementioned chatbots as they are already affecting communities today. Industry researchers estimate that by 2025, globally, cybercrime will net over US$10.5 trillion annually making it 'more profitable than the global trade of all major illegal drugs combined' (Morgan, 2020). We contend that transnational organised cybercrime as a concept will become increasingly relevant to fully understand this

6 A smart legal contract is a 'piece of code stored on a blockchain that self-executes contract terms when certain conditions are met… following a condition-based structure' that iterates with each successive new action (Szabo, 1997).

threatscape. Similarly, how crime and terror form nexuses, and how those collaborations occur (coercion, corruption or mutually beneficial co-operation, for example) will also feature as criminals and terrorists continue to develop 'supplier and customer' relationships in traditional commerce arrangements. The interplay of cybercrime and terrorism already demonstrates a crimes of influence approach, whereby elements of coercion, manipulation and deceit are used to exploit victims and radicalise followers. This is particularly true of different extremist groups, even those with opposed ideologies, who have been observed looking to each other for complementary skillsets. Generative AI is likely to further fuel this dangerous trend towards composite and converging forms of 'salad bar extremism'.

Whether at a CaaS level or via joint ventures, such collaborations depend significantly on the opportunities and risks involved for both parties. Even then, 'such relationships tend to be strongest when criminals and terrorists share a geographic space that provides them with common criminal opportunities'. Such is the case in Afghanistan's poppy trade, Islamist militants' foreign hostage taking and selling throughout the Levant, and Colombian cocaine trafficking in the Sahel (Williams, 2018; White House, n.d.). While these types of crimes do not outwardly appear to be cyber-oriented, they are certainly cyber-enabled. From encrypted communications networks like the now defunct Anom app (DOJ, 2021) to the FBI's shutdown of online drug market 'Silk Road' (FBI, 2015), the way traditional transnational organised criminal groups have adapted to exploit cyber-enabled opportunities, has been well documented. Similarly, the use of the internet to radicalise and recruit susceptible people towards non-violent and violent extremism is also evidenced by a body of significant scholarship (Koehler, 2014; Valentini et al., 2020; Khalil, 2021). How crime-terror nexus types of collaborations manifest to leverage crimes of influence, supported by generative AI-led CaaS, is yet to be seen.

There are also other types of cyber-enabled activities that often fall short of meeting legislative thresholds for action. Such is the case with misinformation and disinformation in so far as freedom of expression and speech in some jurisdictions are, rightly, lawfully protected. In a yet more complex cyber environment, is foreign interference (Davey and Ebner, 2019; Zhang, 2022; Strick, 2023). Another consideration is how these cyber-enabled activities often combine to contribute to an information environment that has the potential to destabilise democracies. Such activities hide in plain sight among conspiracy theories and disinformation campaigns and attempt to influence populations with foreign ideals. While an online contest of ideas is a healthy feature of liberal democracies, the cyber-enabled and often crime-supported environment in which such activities occur are particularly suited to further exploitation by generative AI (Matejic, 2020).

The rapid emergence of generative AI will continue to have major implications for cybercrime and violent extremism, for example. Both have evolved substantially over the past decades in response to emergent technology such as global-positioning-system (GPS), bots, blockchain and more. Existing CaaS products offering services based on these technologies that are able to influence and manipulate opinion, with effects such as election interference and polarisation, have already been observed. For example, Russia's interference in the 2016 United States election has been well documented (US Senate

Select Committee on Intelligence, 2019), as has the use of bots alongside terrorism. After the 2015 ISIS attacks in Paris, for example, the hacker community, Anonymous, claimed to have found and removed 25,000 online bots linked to the terrorist group. The mass manipulation of public opinion through AI has the potential to generate protest, division and instability and to create social conditions within which political violence can easily spiral. Further, the criminal uses of AI, through scams and fraud, extortion, deepfakes, malware and more, provide ample opportunities for terrorists to finance their political activities.

While there is an emergent evidence base to support generative AI-led CaaS being explored by terrorists, it is important not to exaggerate the potential for bad-faith actors to deploy it. As with all criminal endeavours, capability, intent and opportunity remain essential. Similarly, because of its current-state unpredictability, generative AI could also prove to be something of a wildcard for CaaS developers and their criminal or terroristic clientele. Unintended consequences have been observed within AI foundation models by researchers who argue that these 'failure modes' are not yet broadly understood, able to be repaired nor fully explainable. While much has been written about doomsday AI scenarios, including extinction-level events (Stanford University, 2023) which go beyond the focus of this paper, unanticipated outcomes will almost certainly also occur in generative AI-led CaaS cybercrime environments.

At a Five Eyes summit in October 2023, director of the FBI Christopher Wray indicated that the FBI was already aware of terrorists seeking to use AI to assist in building bombs and hiding their activities from authorities (Sabbagh, 2023). Tech Against Terrorism (2023), in an analysis of over 5,000 pieces of AI-generated content, found within terrorist and violent extremist spaces, concerns including media spawning;[7] automated multilingual translation;[8] the generation of fully synthetic propaganda;[9] variant recycling;[10] personalised propaganda;[11] and subverting moderation.[12] While the report

7 Media spawning is the manipulation of content to evade current network detection capabilities (Tech Against Terrorism, 2023).

8 AI-based automated multilingual translation in the context of terrorism and violent extremism is deployed with the intention of overwhelming social and online network linguistic detection mechanisms, to circumvent content moderation processes (Tech Against Terrorism, 2023).

9 Generative AI-developed fully synthetic media in the context of terrorism and violent extremism includes the production of speeches, videos and interactive environments that are used to spread propaganda, and to radicalise and recruit followers (Tech Against Terrorism, 2023).

10 Variant recycling is a term used to explain the repurposing of old content and propaganda to create new versions. The primary aim in creating new versions is to defeat content moderation systems, such as hash-matching mechanisms, that prevent the upload of terrorist and violent extremist and other illegal content (Tech Against Terrorism, 2023).

11 Personalised propaganda in the context of terrorism and violent extremism is the customisation of messages to target particular people and/or demographics. Generative AI has the ability to generate propaganda to appeal to different audience segments (Tech Against Terrorism, 2023).

12 Subverting moderation is a tactic terrorists and violent extremists use to specifically engineer propaganda in ways that purposefully bypass existing content-moderation detection mechanisms. This often enables them to post illegal material online that remains accessible for longer periods due to the time it takes community reports to trigger automated and manual content-moderation processes (Tech Against Terrorism, 2023).

concludes that there is a low risk of widespread extremist adoption in the near future, this experimentation indicates an emerging risk where unintended consequential harms may still result. Present-day examples of terrorist exploitation of generative AI include extreme right-wing users creating antisemitic and racist images; a user generating and sharing instructions on how to create memes and propaganda; the production of a security guide by an Islamic State supporter; another Islamic State supporter using speech-recognition systems to transcribe and translate leadership speeches; al Qaeda supporters producing propaganda; and in the recent attack on Israel by Hamas, Izzd Ad-din Al-Qassam Brigades have used generative AI to produce a small amount of synthetic content to 'augment narrative appeal'. A significant proportion of the posts studied incited violence against Jewish, Black and other minority communities (Tech Against Terrorism, 2023). Cybercriminals and organised crime syndicates have also been observed adopting CaaS generative AI tools. FraudGPT,[13] has been deployed to deliver 'highly convincing phishing emails and deceptive websites' for US$200 a month or an annual subscription of US$1,700. Similarly, WormGPT has been 'purpose built' to 'craft compelling personalised emails'. Researchers have observed WormGPT exhibiting 'astute and persuasive' capabilities. Another new CaaS-led generative AI product is pretexting – a social-engineering tool that 'fabricates stories or pretexts to deceive users into divulging sensitive information.' Researchers have noted a significant rise in the instances of pretexting during social-engineering exploits because it is highly effective in mimicking the writing styles, languages and linguistic proficiency of the real-world organisations they pretend to be (Falade, 2023).

Based on known current-state exploration and use of generative AI by nefarious actors, allied behaviours and their likely outcomes could include the artificial representations of violent atrocities, for example, increasing the potential of such content to provoke extremists to take violent action against civilians in retaliation. Similarly, generative AI has a great deal of potential to exploit the suggestibility of people, such as isolated and disgruntled young men. This is particularly the case with younger people, a demographic of increasing importance in extremism. Teenagers are often still forming their identities, are highly impressionable, and are more likely to be impulsive, taking risks to impress. Generative AI holds substantial potential to radicalise susceptible youth. Even relatively simple chatbots can play a key role in radicalising individuals to violence. In July 2023 a young man pleaded guilty to planning to assassinate the late Queen of the United Kingdom and Commonwealth after encouragement by a chatbot which he had created on the Replika app. A psychiatrist found that the man had 'formed an emotional and sexual relationship' with the bot over months of online interaction.[14] AI chat and other tools have also been observed to provide logistical expertise for potential terrorists or members

13 FraudGPT is a subscription-based tool that, while based on ChatGPT, has been designed to deliver fake content at scale to dupe victims into believing they are dealing with a trusted institution (Falade, 2023).

14 https://www.theguardian.com/uk-news/2023/jul/06/ai-chatbot-encouraged-man-who-planned-to-kill-queen-court-told

of terrorist organisations by providing information on bomb or weapon making, how to behead someone, and how to avoid detection (Lakomy, 2023; Bunker and Bunker, 2023). These activities are all influence-dependent, relying on the participation of susceptible people (whether by ignorance, choice, coercion or deception) to conduct harm.

That current-state generative AI CaaS models are unlikely to face much resistance and are likely to continue to evolve unimpeded, at least over the short term, is of concern. However, the newness of generative AI technology also reduces the likelihood that those tasked with preventing and investigating cybercrime and terrorist activity will be adequately prepared to deal with such technically complex and resource-intensive tasks. While criminals and terrorists can simply purchase a CaaS product online at any time, law enforcement and intelligence agencies are rightly beholden to a range of legislative checks and balances to procure and deploy even the smallest of counter-capabilities. Where capability does exist, legislative remits may not. The result will be the creation of a permissive space for crimes to be committed and terrorist plots to succeed.

Future threatscape

By taking a horizon-scanning approach to future generative AI-led cybercrime alongside a maturing understanding of present-day CaaS innovation, researchers and scholars have begun forecasting risks. HYAS, a company which specialises in cyber-adversary infrastructure, explains that while AI and generative AI are already being exploited by cybercriminals who are creating increasingly sophisticated autonomous cognitive threat agents, generative AI tools have 'the potential to completely revolutionise the landscape of cyber threats (because they) mimic the adaptability of biological viruses, constantly observing their environment and mutating to exploit beneficial circumstances... an opportunistic predator... who can choose its targets and decide when to lay dormant and how to strike to maximise its impact' making it an ever-present, dynamic threat. At a forensic level, polymorphic malware can also 'alter its appearance and behaviour' all while continuing to seek out targets and executing exploits. This technological evolution speaks to an increasingly sophisticated social-engineering capability (HYAS, 2023) and indication of the complexity of the challenges ahead. While individuals will likely remain targets, so too will entire computer systems of value. The ability to stage an intrusion into systems to cause stock market calamities, currency fluctuations, trade-based crises and widespread financial system chaos (Caldwell et al., 2020) leaves insider threats and poor cybersecurity hygiene in workforces at risk of exploitation. These risks will increase as more organisations adopt AI and generative AI as part of their operations, making the use of AI and generative AI for cybersecurity purposes essential.

The exploitation of biases by social engineers also presents significant risks to social cohesion, geopolitical stability and democracy. While present-day observations about the harms fake news, misinformation and disinformation are becoming increasingly clear, MI5 and the FBI were astute in their recognition of election meddling and foreign interference as key benefactors from advances in generative AI (Corera, 2023). For

example, New Zealand researchers at the University of Auckland's Hate and Extremism Insights Aotearoa (HEIA) research laboratory observed in a recent report that democratic backsliding is both a cause and effect of disinformation (HEIA, 2023). With disinformation being a well-used conduit for foreign interference and political deepfakes having already surfaced (Appel and Prietzel, 2022), the potential for generative AI-produced content to mislead and deceive is clear. It is, therefore, likely that generative AI-based deception will have long-term, cascading effects on cognition by reshaping people's worldviews on important political and social issues of the day.

To a degree, cybercriminals are already somewhat successful at this, spoofing branded content, for example, and engaging in social-engineering exploits that deliver some results. The future of these types of generative AI-led CaaS, and the impact they will have on victims however, has the potential to do far more than defraud. Early studies into the pervasiveness of augmented and virtual reality devices, for example, have already found that such immersive environments awaken spatial memory (Heller, 2021) which could have cascading effects on the psychological wellbeing of those who become victims of generative AI-led cybercrime. That most victims will have been influenced into co-opting themselves into a process that causes their own harm and distress is of significant concern. Researchers may find some answers to these challenges from within the criminology and psychology disciplines, where the effects of coercive control and intimate partner abuse have been well studied.

Old crimes, new tools, slow legislature

The evolution of generative AI occurs at a time when the legislation available to combat cybercrime, particularly at a transnational organised level, remains globally largely unfit for purpose. While individual nations may have varying degrees of legislative and regulatory options available to apply to cybercrime challenges, in many cases those mechanisms have not kept pace with advances in internet-enabled technology nor the types of behaviours offenders employ to influence others towards harm. From a victim perspective, law enforcement agencies are often ill-equipped to deal with the types and volume of harm being committed online. Even if they were adequately resourced and supported by up-to-date legislative frameworks, such work would fall to an even smaller forensically capable workforce and then an overwhelmed judiciary. Due to the slow-moving nature of legislative change, generative AI will enable crimes of influence to outpace existing capabilities further. These longstanding issues notwithstanding, governments are giving significant thought and committing increasing resources to the challenge. A globally comprehensive AI legislative tracker, which is regularly updated, can be found online courtesy of the International Association of Privacy Professionals.[15]

15 The International Association of Privacy Professionals online 'Global AI Legislation Tracker' (IAPP, 2023) 'identifies legislative policy and related developments' in Australia, Brazil, Canada, China, the European Union, India, Israel, Japan, New Zealand, Saudi Arabia, Singapore, South Korea, the United Arab Emirates, the United Kingdom and the United States. While not globally comprehensive, it does provide a useful (and regularly updated) central repository of knowledge.

In addition to legislative mechanisms, governments, industry, academia and civil society have developed a range of fora that acknowledges not only the globalised nature of the challenge as a whole, but also the critical need for a globally cohesive solution. While some fora see many nations working together on these issues, many of the most successful fora exist within multistakeholder settings that focus on initiatives that co-exist alongside the legislative and regulatory environment. For the purposes of this paper, a brief overview of some of the larger governmental collaborations and successful multistakeholder initiatives that specifically address cybercrime or particular harm types follows.

Government initiatives

The United Kingdom's Bletchley Declaration (2023), taking a broad approach that recognises 'the potential for unforeseen risks stemming from the capability to manipulate content or generate deceptive content' among other cyber concerns, seeks to ensure frontier AI capabilities are safe and that an understanding of concerns and risks is evidence-based (GOV.UK, 2023). The G7 Hiroshima Process on generative AI, while proposing a broad set of guiding principles for organisations developing advanced AI systems, specifically noted that technology development should not develop or deploy 'systems in a way that undermine democratic values, are particularly harmful... facilitates terrorism, enable criminal misuse, or pose substantial risks to safety, security and human rights'. The G7 also committed to multistakeholder work on monitoring tools and accountability (OECD, 2023). The European Union's (EU) AI Act is set for final approval from the European Parliament in April 2024. If the legislation passes as it stands, the Act will become law in 2026 and effectively regulate AI models based on potential risk. The EU AI Act places particular focus on the development of safe AI that upholds EU values and respects human rights. In practice, this approach attempts to deal with the issues that general-purpose AI models have (such as unpredictable uses) while allowing for lower-risk AI innovation (Gibney, 2024). The United States of America's executive order on 'Safe, Secure, and Trustworthy AI' (2023) sets new standards for safety and security to manage the risks of AI. Taking a 'responsible innovation' approach, the order builds on the existing voluntary commitments involving 15 companies that are at the leading edge of AI development. The order specifically directs that protections from 'AI-enabled fraud and deception' are necessary for detecting AI-generated content and authenticating official content' (White House, 2023). Other nations currently exploring legislative and regulatory mechanisms for AI broadly include Australia where a regulatory approach is undergoing public consultation; China which has implemented interim measures to manage generative AI; France which is investigating data-protection breaches; Ireland with a view that generative AI needs to be regulated; Israel which is undertaking public consultation on a national AI policy; Italy which temporarily banned ChatGPT in April 2023 over data-protection concerns; Japan which is seeking to introduce a regulatory framework; and Spain which is also investigating data-protection breaches (Reuters, 2023). The United Nations Educational, Scientific and Cultural Organization (UNESCO)

has called on governments to regulate generative AI in education and research. UNESCO is particularly concerned with the 'harm and prejudice' generative AI may incubate, and is advocating for a global standardised approach (UNESCO, 2023). The United Nations Treaty on Cybercrime is also considering, broadly, cyber-enabled crime. While final negotiations were due to conclude in February 2024, concerns from civil society and industry remain about the treaty's impact on human rights, along with reservations from some nations about the document's scope. The treaty's applicability to generative AI appears to be implied as an enabling technology (Kazakova et al., 2023).

Other legislative instruments with a focus on multijurisdictional cybercrime more broadly include the Budapest Convention, led by the Council of Europe, which is currently the only international treaty that 'criminalises conducts and typologies committed through computer and information systems'. Its main objective as an instrument is to enable seamless information sharing between signatories, providing a legal basis for disclosure and preservation of information across a range of user, subscriber and other forensic data points. With a focus on international co-operation, currently 66 nations are party to the convention. The Cybercrime Convention Committee (T-CY), formed by nations party to the convention, have given some thought to AI as a vector of crime given the neutrality of its drafting 'precisely because the original drafters... anticipated how the threat landscape... would likely evolve'. The Lanzarote Convention, another Council of Europe international treaty, has 48 state signatories but is yet to fully explore how it may be applied in practice within a generative AI cybercrime setting. The Istanbul Convention, another Council of Europe initiative, has 34 state parties and a reference group of experts. Yet the convention itself does not address crimes committed online, although its reference group has made recommendations to address this (Velasco,2022).

Multistakeholder initiatives

There are several multistakeholder initiatives working closely on challenges that are likely to occur as cybercrime and generative AI converge or which are harm-adjacent. New Zealand and France's Christchurch Call, with its focus on eliminating terrorist and violent extremist content online, is working with its multistakeholder community to 'contribute to the development of frameworks... to identify, report, and mitigate terrorist and violent extremist exploitation' of generative AI tools (Christchurch Call, 2023). The Global Partnership on AI (GPAI) with a secretariat hosted by the OECD AI policy observatory, has established a working group on harmful online content such as hate speech. Taking a content-moderation view of potential AI solutions, the GPAI advocates for responsible development and adoption of AI (GPAI, 2023). The illegality of hate speech differs significantly between jurisdictions, which AI and generative AI may be particularly well-suited to navigating in a borderless social media and online environment.

Beyond foreign interference, misinformation and disinformation, thinking on influence as a vector that underpins cybercrime, appears to have been considered in depth only within multistakeholder initiatives. The Christchurch Call, for example, considers the role

of algorithmic influence in radicalisation towards violent extremism. Similarly, the GPAI has conducted research into AI-based online recommender systems in the context of driving users towards terrorism and violent extremism content (Christchurch Call, 2023; GPAI, 2023). These types of initiatives, unlike governmental focus on legislative and regulatory responses, go some way towards building an understanding of the upstream types of influences that users are routinely subjected to that contribute to the creation of downstream victims of cybercrime. However, such knowledge is useful only if its findings are put into practice. The solution here also leans towards multistakeholder initiatives, that are, by design, more agile and able to cultivate cross-sector relationships in a less constrained environment. That is not to diminish the important work of governments, only to point out that the roles and duties citizens expect of their elected officials necessitate a different way of working with their stakeholders.

Conclusion

The cognitive frontier that generative AI heralds has the potential to enable cybercrime in ways that are largely unavoidable for victims. While the technology involved in building environments to deceive, influence and manipulate assist this process, at a fundamental level, generative AI challenges the way human cognition moves between System 1 and System 2 thinking. By quickly building rapport and familiarity with targets, whether by email, text, social media or virtual environments, generative AI's pervasiveness encourages an overreliance on System 1 thinking to avoid the scrutiny System 2 entails.

When considering how generative AI-enabled cybercrime, transnational organised crime and violent extremism may evolve, it is easy to catastrophise a possible future without giving due consideration to the opportunities that will also arise to mitigate, and perhaps even prevent, some of those risks. At a fundamental level, as has been the case with web 2.0, there will always be those who exploit advances in technology for nefarious purposes. This is the nature of humanity, not necessarily the nature of technology. Catastrophising generative AI fixes attention on threats instead of encouraging people to see their way through complexity towards solutions. That generative AI comes at a time of heightened global polarisation, regional conflict, a highly contested information environment and democratic backsliding should not go unnoticed. When faced with change, humans instinctively and consistently default to fearing the unknown. Becoming comfortable with complexity, and daring to try to make sense of it, is a type of mental gymnastics that most find uncomfortable. However, if scholars, researchers, civil society and industry don't push past this discomfort, we risk missing the ground-floor opportunities available to address some of generative AI's biggest risks. Multistakeholder initiatives are well placed to help navigate this environment and any dystopian view of where generative AI may lead humanity needs to be balanced with technological optimism.

There is no doubt that technological guardrails for generative AI will be required. However, any approach to designing these guardrails must be robustly evidence-based. Likewise, the design of effective guardrails is not just a problem for the technology industry.

Robust, defensive protective and detection-capable mechanisms are the combined result of appropriate, and human rights affirming, safety by design frameworks that are underpinned by internationally cohesive legislative conventions supported by reasonable jurisdictional regulation. Without a level of international and virtual legislative cohesion and co-operation, questions such as who will be responsible for generative AI's conduct, who is in charge of investigations of the cybercrimes that arise, and who is responsible for prosecuting such multi-jurisdictional crimes, will remain unanswered. The real-world consequences of failing to approach this challenge with a global outlook is that it will create permissive environments ripe for exploitation.

What is largely missing from this discussion, however, is consideration of the cognitive impact on enabling near-horizon and future cybercrime. Influence rarely occurs in isolation yet thrives in vacuums. Without addressing the psychosocial aspects of generative AI-led cybercrime, work in this domain will only focus on 'bottom-of-the-cliff' solutions that come too late to be considered preventative. While, in part, this can be attributed to the emergent nature of both the type of offending and understanding of this field of study, an overreliance on risk as a lens to inform expected victim outcomes is problematic. By taking a speculative approach to conceptualising risk in this context, in isolation of understanding real-world as-it-happens victim experiences, researchers miss qualified insight into how these exploits were able to be detected, what kinds of persuasive behaviours were employed, whether the exploits succeeded or failed, and why. By studying victim experiences in the generative AI-led CaaS environment, scholars and practitioners can begin to develop protective and preventative solutions to address the generative AI-led crimes of influence of the near future.

Conflict of interest

The authors declare that they have no conflicts of interest. This paper does not represent the views or policy of any of the authors' employers, associations or affiliations.

Acknowledgments

Thanks to associate professor Nick O'Brien for feedback on an earlier version of this paper.

References

Anthropic (2023) Claude chatbot. Query: 'Tell me how generative AI will change cybercrime'. Accessed and response provided on 14 October 2023. A full transcript of the query and response is available on request.

Appel, M. and F. Prietzel (2023) The Detection of Political Deepfakes. *Journal of Computer-Mediated Communication*, 27(4) pp. 1–13.

Ariely, D. (2008) *Predictably Irrational*. New York: Harper Collins.

Bunker, R.J. and K.O.K. Bunker (2023) *The Terrorism Potentials of ChatGPT and Related Generative AI Models*. A C/O Futures Terrorism Research Note Series.

Byrne, M.D. (2023) Generative Artificial Intelligence and ChatGPT. *Journal of PeriAnesthesia Nursing*, 38 pp. 519–522.

Caldwell, M., J.T.A. Andrews, T. Tanay and L.D. Griffin (2020) AI-enabled future crime. *Crime Science*, 9(14) pp. 1–13.

ChatGPT (2023) an OpenAI chatbot. Query: 'Tell me how generative AI will change cybercrime'. Accessed and response provided on 14 October 2023. A full transcript of the query and response is available on request.

Chebac, A. (2023) *What is Cybercrime-as-a-Services (CaaS)?* Heimdal. Accessed on 25 October 2023 at: https://heimdalsecurity.com/blog/what-is-cybercrime-as-a-service-caas/

Christchurch Call (2023) *2023 Leader's Summit Joint Statement*. Accessed online 11 November 2023 at: https://www.christchurchcall.com/assets/Documents/Christchurch-Call-Leaders-Summit-2023-Joint-Statement-ENG.pdf

Cialdini, R. (2021) *Influence. The Psychology of Persuasion*. New and expanded edition. Harper Collins. pp. 23–363.

Cialdini, R. (2016) *Pre-suasion*. Random House. pp. 7–8.

Corera, G. (2023) AI risks are unknown even to GCHW, Anne Keast-Butler tells BBC. BBC. Accessed on 5 November 2023 at: https://www.bbc.com/news/uk-67301402

Cornell Law School (2024) Legal Information Institute. *Crime of Passion*. Accessed on 17 February 2024 at: https://www.law.cornell.edu/wex/crime_of_passion

Davey, J. and J. Ebner (2019) *The Great Replacement: The Violent Consequences of Mainstream Extremism*. Institute of Strategic Dialogue.

Department of Justice (DOJ), United States of America (2021) *FBI's Encrypted Phone Platform Infiltrated Hundreds of Criminal Syndicates; Result is Massive Worldwide Takedown*. Accessed on 16 November 2023 at: https://www.justice.gov/usao-sdca/pr/fbi-s-encrypted-phone-platform-infiltrated-hundreds-criminal-syndicates-result-massive

Enoch, J., L. McDonald, L. Jones, P.R. Jones and D.P. Crabb (2019) Evaluating Whether Sight Is the Most Valued Sense. *JAMA Opthalmol*, 137(11), pp. 1317–1320.

Falade, P.V. (2023) Decoding the Threat Landscape: ChatGPT, FraudGPT and WormGPT in Social Engineering Attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,* (9)5, pp. 185–198.

Federal Bureau of Investigation (FBI), United States of America (2015) *Ross Ulbricht, the Creator and Owner of the Silk Road Website, Found Guilty in Manhattan Federal Court on All Counts*. Accessed on 16 November 2023 at: https://www.fbi.gov/contact-us/field-offices/newyork/news/press-releases/ross-ulbricht-the-creator-and-owner-of-the-silk-road-website-found-guilty-in-manhattan-federal-court-on-all-counts

Gibney, E. (2024) What the EU's Tough AI Law Means for Research and ChatGPT. *Nature Articles*. Accessed on 17 February 2024 at: https://doi.org/10.1038/d41586-024-00497-8

Global Partnership on Artificial Intelligence (GPAI) (2023) *What we do*. Accessed on 27 October 2023 at: https://www.gpai.ai/projects/

Google Gemini (2023) a Google chatbot. Query: 'Tell me how generative AI will change cybercrime'. Accessed and response provided on 14 October 2023. A full transcript of the query and response is available on request.

GOV.UK (2023) *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023*. Policy Paper. Accessed on 3 November 2023 at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Harms, M.B. (2017) Stress and Exploitative Decision-making. *Journal of Neuroscience*, 37(42), pp. 10035–10037.

Hartvigsson, T. (2021) Between Punishment and Care: Autonomous Offenders Who Commit Crimes Under the Influence of Mental Disorder. *Criminal Law and Philosophy*, 17, pp. 111–134.

Hate and Extremism Insights Aotearoa (HEIA) (2023) *Disinformation Trends in New Zealand: A HEIA Snapshot Report*. 20. Accessed online 2 January 2024 at: https://www.heiaglobal.com/post/disinformation-trends-in-new-zealand

Heller, B. (2021) Watching Androids Dream of Electric Sheep: Immersive Technology, Biometric Psychography, and the Law. *Vanderbilt Law Review*, 23(1). https://scholarship.law.vanderbilt.edu/jetlaw/vol23/iss1/1

Hendrycks, D., M. Mazeika and T. Woodside (2023) An Overview of Catastrophic AI Risks. Version 6, 9 October 2023.

Hong Kong Computer Emergency Response Team Coordination Centre (HKCERT) (2023) *Unmasking Cybercrime-as-a-Service: The Dark Side of Digital Convenience*. Accessed on 24 October 2023 at: https://www.hkcert.org/blog/unmasking-cybercrime-as-a-service-the-dark-side-of-digital-convenience

HYAS (2023) *EyeSpy: Proof of Concept*. Accessed on 26 October 2023 at: https://www.hyas.com/blog/eyespy-proof-of-concept

IBM (2023) What's Next in AI is Foundation Models at Scale. Accessed on 13 October 2023 at: https://research.ibm.com/artificial-intelligence

International Association of Privacy Professionals (IAPP) (2023) *Global AI Legislation Tracker*. Accessed on 28 October 2023 at: https://iapp.org/resources/article/global-ai-legislation-tracker/

Kahneman, D. (2011) *Thinking, Fast and Slow*. New York: Farrar, Strauss and Giroux. pp. 20–21; 97–99.

Kazakova, A., K. Swift and B. Kovač (2023) *Key Takeaways from the Sixth UN Session on Cybercrime Treaty Negotiations*. Geneva Internet Platform, Digwatch. Accessed on 25 November 2023 at: https://dig.watch/updates/key-takeaways-from-the-sixth-un-session-on-cybercrime-treaty-negotiations

Keeler, J.D. (2024) Bribery and Corruption: Crimes of Influence. *Law N Guilt*. Accessed on 17 February 204 at: https://www.lawnguilt.com/bribery-and-corruption-crimes-of-influence/

Khalil, L. (2021) *GNET Survey on the Role of Technology in Violent Extremism and the State of Research Community – Tech Industry Engagement*. Global Network on Extremism and Technology.

Koehler, D. (2014) The Radical Online: Individual Radicalization Processes and the Role of the Internet. *Journal for Deradicalization*, 15(1), pp. 116–134.

Lakomy (2023) Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impacts of Chatbots and Image Generators on Online Terrorist Activities. *Studies in Conflict & Terrorism*. DOI: 10.1080/1057610X.2023.2259195

Lewis, J. (2018) *Economic Impact of Cybercrime – No Slowing Down*. A joint McAfee and Center for Strategic and International Studies (CSIS) Report. Accessed on 29 October 2023 at: https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/economic-impact-cybercrime.pdf

Makarenko, T. (2004) The Crime-Terror Continuum: Tracing the Interplay between Transnational Organised Crime and Terrorism. *Global Crime*, 6(1), pp. 129–145.

Matejic, N. (2020) *2040: An Information Odyssey*. NATO Innovation Hub: Warfighting in 2040 Report.

Meta (2021) *Trademark/Service Mark Application*. Principal Register. Serial Number: 97097362. Filed on 28 October 2021. Accessed on 1 November 2023 at: https://tsdr.uspto.gov/documentviewer?caseId=sn97097363&docId=APP20211101091335&linkId=9#docIndex=8&page=1

Morgan, S. (2020, November 13) Cybercrime to Cost the World $10.5 Trillion Annually by 2025. *Cybercrime Magazine*. Accessed on 19 November 2023 at: https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/

Munn, L. (2020) Angry by Design: Toxic Communication and Technical Architectures. *Humanities and Social Sciences Communications*, 7, 53. https://doi.org/10.1057/s41599-020-00550-7

Nickerson, R.S. (1998) Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2, pp. 175–220.

Nvidia (2023) *What is Generative AI?* Accessed on 12 October 2023 at: https://www.nvidia.com/en-us/glossary/data-science/generative-ai/

Organisation for Economic Co-operation and Development (OECD) (2023) *G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a Common Understanding of Generative AI*. Report prepared for the 2023 Japanese G7 Presidency and the G7 Digital and Tech Working Group. Accessed on 1 October 2023 at: https://www.oecd.org/publications/g7-hiroshima-process-on-generative-artificial-intelligence-ai-bf3c0c60-en.htm

Reuters (2023, September 11) What are Governments Doing to Try to Regulate AI? *euronews.next*. Accessed on 25 November 2023 at: https://www.euronews.com/next/2023/09/11/which-countries-are-trying-to-regulate-artificial-intelligence

Sabbagh, D. (2023, October 18) Terrorists Could Try to Exploit Artificial Intelligence, MI5 and FBI Chiefs Warn. *The Guardian*. Accessed on 23 October 2023 at: https://www.theguardian.com/technology/2023/oct/18/terrorists-exploit-artificial-intelligence-ai-mi5-fbi-chiefs-warn

Simi, P., K. Blee, M. DeMichele and S. Windisch (2017) Addicted to Hate: Identity *Residual* among Former White Supremacists. *American Sociological Review*, 82(6), pp. 1167–1187.

Stanford University (2023) *The Stanford Emergency Technology Review 2023: A Report on Ten Key Technologies and their Policy Implications*. pp. 21–29.

Strick, B. (2023) *Incitement to Kill: Tracking Hate Speech Targeting Ukrainians During Russia's War in Ukraine*. Centre for Information Resilience.

Synnott, A. (2022) Sight Is Our Most Dominant Sense, But Is It Trustworthy? *Psychology Today*. Accessed on 17 February 2024 at: https://www.psychologytoday.com/intl/blog/rethinking-men/202207/sight-is-our-dominant-sense-is-it-trustworthy

Szabo, N. (1997) The Idea of Smart Contracts. Accessed online 18 February 2024 at: https://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/idea.html

Tech Against Terrorism (2023, November 8) *Early terrorist experimentation with generative artificial intelligence services*. Briefing. Accessed on 10 November 2023 at: https://techagainstterrorism.org/news/early-terrorist-adoption-of-generative-ai

Thaler, R. and C. Sunstein (2009) *Nudge. Improving Decisions about Health, Wealth and Happiness*. Penguin Books. pp. 19–112.

United Nations Educational, Scientific and Cultural Organization (UNESCO) (2023) UNESCO calls for regulations on AI use in schools. Accessed on 17 October 2023 at: https://news.un.org/en/story/2023/09/1140477#:~:text=The%20UN%20Educational%2C%20Scientific%20and,data%20protection%20and%20user%20privacy.

UOLLB First Class Law Notes (2024) *Under Influence in Contract and Criminal Law*. Accessed on 17 February 2024 at: https://uollb.com/blog/law/under-influence-in-contract-and-criminal-law

US Senate Select Committee on Intelligence (2019) *Senate Intel Releases Election Security Findings in First Volume of Bipartisan Russia Report*. Accessed on 21 November 2023 at: https://www.intelligence.senate.gov/press/senate-intel-releases-election-security-findings-first-volume-bipartisan-russia-report

Valentini, D., A.M. Lorusso and A. Stephan (2020) Onlife Extremism: Dynamic Integration of Digital and Physical Spaces in Radicalization. Hypothesis and Theory Article. *Frontiers in Psychology*, 11, 524. https://doi.org/10.3389/fpsyg.2020.00524

Velasco, C. (2022) Cybercrime and Artificial Intelligence. An Overview of the Work of International Organizations on Criminal Justice and the International Applicable Instruments. *ERA Forum*, 23, pp. 109–126. https://doi.org/10.1007/s12027-022-00702-z

Wanless, A. and M. Berk (2017) Participatory Propaganda: The Engagement of Audiences in the Spread of Persuasive Communications. Paper delivered at Social Media and Social Order conference, November/December 2017, Oslo, Norway.

The White House (2023) FACT SHEET: President Biden Issues Executive Order on Safe, Secure and Trustworthy Artificial Intelligence. Accessed online 1 November 2023 at: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

The White House, President Barack Obama, National Security Council (n.d) Transnational Organized Crime: A Growing Threat to National and International Security. Accessed on 19 November 2023 at: https://obamawhitehouse.archives.gov/administration/eop/nsc/transnational-crime/threat

Williams, P. (2018) The Organized Crime and Terrorist Nexus: Overhyping the Relationship. Stratfor Worldview.

Wolbers, R. (2023) *The Future of the Christchurch Call to Action. How to Build Multistakeholder Initiatives to Address Content Moderation Challenges*. Ian Axford (New Zealand) Fellowships in Public Policy.

World Economic Forum (2023, February 6) *Artificial Intelligence. What is generative AI? An AI explains*. Accessed on 12 October 2023 at: https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/

Zak, P.J. (2017) The Neuroscience of Trust. *Harvard Business Review*. pp. 84–90. https://hbr.org/2017/01/the-neuroscience-of-trust

Zhang, A., T. Hoja and J. Latimore (2022) *Gaming Public Opinion: The CCP's Increasingly Sophisticated Cyber-Enabled Influence Operations*. Australian Strategic Policy Institute International Cyber Policy Centre. https://ad-aspi.s3.ap-southeast-2.amazonaws.com/2023-05/Gaming%20public%20opinion.pdf?VersionId=QYkBIWncbBU0E1KAhg9mX3TD7kwlWcWj

About the authors

Nicole Matejic is a national security-focused behavioural economist and adjunct lecturer at Charles Sturt University in Australia, and completed this paper while also on secondment with the Department of Prime Minister and Cabinet's Christchurch Call Unit in Aotearoa, New Zealand.

Chris Wilson is a senior lecturer in politics and international relations at the University of Auckland (Aotearoa New Zealand) and CEO of Hate and Extremism Insights Aotearoa (HEIA).