Pauwels, Eleonore

Working Paper

Preparing for next-generation information warfare with generative AI

CIGI Papers, No. 310

Provided in Cooperation with:
Centre for International Governance Innovation (CIGI), Waterloo, Ontario

Suggested Citation: Pauwels, Eleonore (2024) : Preparing for next-generation information warfare with generative AI, CIGI Papers, No. 310, Centre for International Governance Innovation (CIGI), Waterloo, ON, Canada

This Version is available at:
https://hdl.handle.net/10419/311791

Centre for International
Governance Innovation

CIGI Paper No. 310 — December 2024

Preparing for Next-Generation Information Warfare with Generative AI

Eleonore Pauwels

Centre for International
Governance Innovation

CIGI Paper No. 310 — December 2024

Preparing for Next-Generation Information Warfare with Generative AI

Eleonore Pauwels

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

Credits

Managing Director and General Counsel Aaron Shull
Director, Program Management Dianna English
Program Manager and Research Associate Kailee Hilt
Senior Publications Editor Jennifer Goyder
Publications Editor Christine Robertson
Graphic Designer Sepideh Shomali

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Table of Contents

About the Author

Eleonore Pauwels is an international expert in the security, societal and governance implications generated by the convergence of artificial intelligence (AI) with other dual-use technologies, including cybersecurity, genomics and neurotechnologies. Eleonore provides expertise to the World Bank, the United Nations and the Global Center on Cooperative Security in New York. She also works closely with governments and private sector actors on AI-Cyberthreats Prevention, the changing nature of conflict, foresight and global security. In 2018 and 2019, she served as research fellow on emerging cybertechnologies for the United Nations University's Centre for Policy Research. At the Woodrow Wilson International Center for Scholars, she spent 10 years within the Science and Technology Innovation Program, leading the Anticipatory Intelligence Lab. She is also part of the Scientific Committee of the International Association for Responsible Research and Innovation in Genome Editing. Eleonore regularly testifies before US and European authorities, including the US Department of State, the National Intelligence Council, the US Bipartisan Commission on Biodefense, the Council of Europe, the European Commission and the United Nations. She writes for *Nature*, *The New York Times*, *The Guardian*, *Scientific American*, *Le Monde*, *Slate*, UN News, *The UN Chronicle* and the World Economic Forum.

Executive Summary

Modern conflicts involve the weaponization of information and the manipulation of human behaviours. Artificial intelligence (AI) and its integration into individuals' daily lives promises to augment, accelerate, but also complicate these trends. Two important shifts will help us understand this emerging warfare for what it truly is: an attack on humanity itself.

AI is making information warfare more powerful and more accessible. Generative AI combined with data capture provides new techniques to industrialize the offensive use of disinformation. In addition, the integration of generative AI with other powerful technologies complexifies the potential of information warfare. What is at stake is the weaponization of dual-use knowledge itself. Generative AI is already learning to democratize military and civilian expertise in technological domains as complex as AI, neuro-, nano- and biotechnology. Such capacity will provide both state and non-state actors with access to knowledge and mentorship related to impactful technologies. This diffusion of power will change the nature of information and physical warfare, increasing dual-use knowledge asymmetries between threat actors in conflicts. There is an urgent need to prepare for misuse scenarios that harness technological convergence. New converging risks will bring collective security challenges that are not well understood or anticipated globally.

Introduction

The world has entered a complex and dangerous decade. As new and old threats converge and challenge the multilateral order, one of the most seismic shifts is taking place at the intersection of war, technology and cyberspace. Modern conflicts — whether they are declared, contested or waged in the grey zone — are amplified by a technological revolution that is inherently dual use. These conflicts merge physical and digital fronts, invading cities and factories, homes and everyday devices, and producing new targets and victims in their wake. Frontiers between peace and war, offence and defence, civilian and military technologies, and state forces and cyberproxies are fading.

Modern conflicts increasingly involve the weaponization of information and the manipulation of human behaviours and perceptions. The rapid development of artificial intelligence (AI) and its integration into individuals' daily lives and societies' inner structures promises to not only augment and accelerate, but also complicate these trends. This paper aims to demonstrate two important shifts that will help us to recognize and understand this emerging warfare for what it truly is: an attack on humanity itself.

First, AI is making information warfare both more powerful and more accessible by acting as a catalyst. The development of generative AI, combined with diverse forms of data capture, provides new techniques to drastically improve, tailor, scale up and even industrialize the offensive use of disinformation. For instance, personalized AI assistants and chatbots now have the capability to engage users in seemingly authentic conversations, subtly injecting manipulative content tailored to the user's psychological profile and preferences. With persuasive narratives, sophisticated bot networks can deeply influence individual and group beliefs. These battles for influence, supercharged by algorithms, are being waged for the control of people's emotions and attitudes and are the prevailing means of undermining social cohesion and trust. In times of conflict, these tools impact critical elements of civilian protection and civilian decision making for survival, causing direct and indirect harms to populations (United Nations General Assembly 2022). For marginalized populations and vulnerable groups, such as women and youth, they may increasingly condition and limit notions of self-determination, and could continue to do so with future generations to come.

Second, the integration of AI, including generative AI, with other powerful technologies is a radical shift as this convergence broadens and complexifies the potential of information warfare. What is at stake is the weaponization of dual-use knowledge itself, including possibly all forms of dual-use expertise developed by human civilization.

Generative AI is already learning to democratize[1] strategic military and civilian expertise and tacit knowledge in technological domains as complex as AI, neuro-, nano- and biotechnology. Such capacity will provide a diversity of both state and non-state actors with access to sensitive knowledge and mentorship related to impactful technologies. This diffusion of power will change not only the scale, but also the nature of both information and physical warfare, increasing dual-use knowledge asymmetries between threat actors involved in conflicts. There is an urgent need to prepare for adversarial use or misuse scenarios that could harness the convergence of what were presented as primarily civilian, beneficial technologies. This new confluence of risks will bring collective security challenges that are not well understood or anticipated globally.

The stakes are high. As AI and generative AI systems reshape how knowledge, expertise and information are used and potentially manipulated in conflict and in the grey zone between war and peace, now is the time to think forward and assess risks, vulnerabilities and forms of resilience. While there will be specific implications for military forces and strategic thinking, prevention and resilience will depend on a whole-of-society response.

The strategic goals of this paper are twofold. First, by providing in-depth analysis on how AI, in convergence with other technologies, can be used to amplify information warfare, the paper aims to inform military authorities and strategic thinkers, policy makers and legal experts, and civil society and multilateral institutions about the emerging strategies at play that have the potential to threaten and weaken societies while escaping accountability. Second, by analyzing how international law applies to emerging forms of information warfare, this paper aims to identify legal gaps and ambiguities as well as potential entry points to support governance and policy processes at national and multilateral levels.

The paper opens with a framing section to define the topic of concern and rapidly review how recent conceptualizations of information warfare have combined with trends related to digital transformation and the evolving conflict landscape (see Box 1). The technical section sheds light on the two major shifts mentioned above, demonstrating how AI not only democratizes information warfare, but also complexifies and broadens its ramifications. The technical section will therefore cover specific uses of AI in information warfare, including psychological operations; the implications for military forces and civilian populations; a few recent real-world manifestations; and AI's future potential and convergence with other technologies. A detailed scenario follows, demonstrating how, at a time of protracted armed conflict, AI-led information warfare could harness dual-use knowledge and sophisticated techniques in biotechnology to undermine public authorities and psychologically destabilize civilian populations. The legal section reviews the protective measures afforded in the international legal framework, as well as the legal gaps and ambiguities that may inhibit effective protection and accountability. This final section closes with highlighting the need for civil-military synergies and elements of a whole-of-society response to strengthen prevention and resilience.

Framing Section

Information warfare is not a new phenomenon, and conflicts of the past have involved deception. This paper does not aim to retrace the history of war propaganda and review extensive literature but instead builds on interviews[2] with technical and legal experts to analyze recent trends and their implications. The goal is to show how the integration of new technologies into our networked world is changing not only the scale, but also the nature and power of war and information warfare.

Transformative shifts are taking place at the intersection of war, technology and cyberspace. To understand how these shifts impact and shape information warfare, we need to look at a series of converging trends. They concern the

1 In this context, the term "democratize" implies that generative AI is helping to spread and share advanced knowledge and capabilities that were previously confined to a select group of experts in the military or specialized civilian fields. This process allows more individuals and organizations to leverage high-level strategic insights and tools, thereby allowing outsourcing to a diversity of actors across different sectors.

2 In 2022–2023, the author conducted a series of interviews with experts in AI and cybersecurity, security implications of emerging technologies, civilian protection in conflict, policy and international law. This research paper builds on the insights, signals and foresight discussed during those interviews.

global revolution that constitutes the digital transformation of our societies; multiple ways to harness information warfare in a multipolar environment; and the evolving nature of conflicts.

The Digital Revolution

The internet has become a laboratory for information warfare, a new theatre of war where information itself is weaponized (United Nations General Assembly 2022). In their excellent monograph, *LikeWar: The Weaponization of Social Media*, Peter Warren Singer and Emerson T. Brooking demonstrate how social media has created a new global environment for conflict, blurring distinctions between civilian and military functions and actions in the digital and physical realms (Singer and Brooking 2018). The rapid digital transformation of our societies has increasingly merged civilian and military technologies, creating new dependencies between the digital architectures that power private, public and national security systems. On this internet battlefield that defies the control of military forces and governments, supremacy is achieved through the command of attention and pervasive forms of psychological and algorithmic influencing. In Singer and Brooking's words, "because virality can overwhelm truth, what is known can be reshaped" (ibid., 22).

Despite its adaptive nature, several critical trends that have characterized information warfare in the past are still relevant today. Back in 2014, the year the Russian Federation annexed Crimea, Peter Pomerantsev and Michael Weiss published *The Menace of Unreality: How the Kremlin Weaponizes Information, Culture and Money*, shedding light on how, at its core, the Russian system of information manipulation relies on nihilism about the existence of objective truth (Pomerantsev and Weiss 2014). Reminiscent of what William Hutchinson wrote nearly two decades ago, we realize that, in modern conflicts, information can be both "targeted" and "weaponized" and that psychological influence has become ever more critical to control populations at home and across borders through global reach (Hutchinson 2006). In the near future, the advent of immersive digital spaces could amplify the ways that every one of us is involved in modern conflicts, with the potential to blur even further the line between reality and deception and to mobilize large swaths of populations, resources and weapons around deceiving narratives.

A Multipolar Cyberspace

In the current multipolar geopolitical landscape, authoritarian regimes such as Russia, China and Iran have relied on proxies and increasingly employed sophisticated information warfare tactics to advance their strategic interests. Leveraging disinformation, these states aim to undermine public trust in democratic institutions, manipulate global narratives and destabilize their adversaries. The COVID-19 pandemic and subsequent vaccination efforts provided fertile ground for these information operations, highlighting their capability to exploit crises for geopolitical gains. The confluence of information warfare and biological threats is particularly relevant to this paper's scenario.

Russia has a long history of using disinformation as a tool of statecraft, aiming to create confusion and weaken the resolve of its adversaries. During the COVID-19 pandemic, Russian actors disseminated false information about the origins and impacts of the virus. State-controlled media and proxy websites promoted conspiracy theories, suggesting that the virus was a bioweapon developed by the United States (Mouton, Lucas and Guest 2023; Moy and Gradon 2023). This narrative was designed to sow distrust and exacerbate tensions between the United States and its allies. In addition, Russia's disinformation efforts targeted the vaccination campaigns of Western countries. Russian media spread misleading information about the safety and efficacy of Western vaccines such as Pfizer and Moderna, while promoting its own Sputnik V vaccine as a superior alternative (Schafer et al. 2021). These efforts aimed to undermine public confidence in Western vaccines, thereby slowing vaccination rates and prolonging the pandemic's impact on Western societies (MacDonald and Ratcliffe 2023; Mouton, Lucas and Guest 2023; Whiskeyman and Berger 2021).

China has also been at the forefront of combining information warfare with AI innovation to influence global perceptions and advance its strategic objectives (Beauchamp-Mustafaga 2024). During the pandemic, Chinese state media and online platforms disseminated positive narratives about China's handling of the outbreak, contrasting its success with the perceived failures of Western countries (MacDonald and Ratcliffe 2023; Beauchamp-Mustafaga 2024; Moy and Gradon 2023; Whiskeyman and Berger 2021). This was part of a broader strategy to deflect blame and position China as a global leader in pandemic management. China's disinformation campaigns

extended to vaccine diplomacy. Chinese state media cast doubt on the safety and effectiveness of Western vaccines, while promoting Chinese-made vaccines such as Sinovac and Sinopharm (Schafer et al. 2021). This narrative was aimed at enhancing China's soft power and expanding its influence in regions such as Africa, Latin America and Southeast Asia, where vaccine diplomacy could translate into geopolitical leverage.

Iran has utilized information warfare to target both regional rivals and the broader international community. Iranian state media circulated conspiracy theories suggesting that the COVID-19 virus was an American biological weapon, aiming to stoke anti-American sentiment and divert attention from Iran's own domestic challenges (Mouton, Lucas and Guest 2023). Iranian disinformation also targeted vaccination efforts: false information about the dangers of Western vaccines was spread by Iranian media, contributing to vaccine hesitancy and undermining public health efforts (Schafer et al. 2021; Whiskeyman and Berger 2021). This strategy was part of a larger effort to portray Iran as resilient and self-sufficient, capable of managing the pandemic without Western assistance (MacDonald and Ratcliffe 2023; Mouton, Lucas and Guest 2023).

What we also see emerging is how nation-states are increasingly engaging in various types of knowledge and technological transfer (such as AI) with proxy actors to target civilian populations alongside traditional political, economic and information aspects of advanced geopolitical conflicts. As mentioned by Wesley R. Moy and Kacper T. Gradon, "from reconnaissance activities and the profiling of target audiences to the weaponization of distorted or fake information and psychological operations, AI broadens the potential of information operations" (Moy and Gradon 2023, 57).

The Evolving Nature of Conflicts

Civilian populations are increasingly targeted in wartime disinformation and suffer enduring harms. Manipulation of information in conflict is increasingly used to legitimate direct acts of violence against civilians and recruit youth into offensive operations, leading to highly traumatic physical and mental harms (Katz 2021; United Nations General Assembly 2022). Another disturbing form of adversarial manipulation harnessed by parties to conflict is the distortion of information that is related to humanitarian and medical efforts and vital to secure human needs (Katz 2021; Morris 2024). The brunt of these evolving practices of information warfare have been suffered by civilian populations, with a particularly vivid impact on women and children (United Nations General Assembly 2022; see Box 3).

What we see materializing before our eyes is a polymorphous type of warfare that merges cyberattacks and information operations and is waged by states or their proxies, sometimes in hostile situations that do not clearly meet the legal threshold of an armed conflict, in the "grey zone" between war and peace (Pauwels 2024). We have entered a new era of hybrid warfare where non-military tactics coexist or are coordinated with kinetic warfare, and target both military forces and civilians (Burt 2023; Khan 2023). There is also an increasing risk of facing a privatization of information warfare. The world has been forced to cope with surrogates and mercenaries in the past, when outsourcing war depended largely on arms trade and trafficking; proxies today, however, thrive on the intangible transfer of dual-use knowledge and democratized access to related technologies. Recent research has pointed to the harmful merger between two growing industries — those that trade cyberweapons and those that industrialize cybercrime — and the offensive proxy capacities these industries bring to an increasing number of nation states and violent actors (Pauwels 2024). Nation-states and their proxies have the potential to harness the integration of AI and converging technologies in information warfare and pose new systemic risks in conflicts.

Box 1: Primary Strategic Terms and Scope

There is no internationally agreed conceptualization of what constitutes information warfare and how it manifests in armed conflict. Top-down definitions of information warfare vary between tech-leading nations such as the United States, Russia and China. Even within a national context, military institutions and doctrines might not necessarily share the exact same concepts and scope. For instance, the US Navy approaches information as purely raw data or digital signals, while the US Department of Defense also considers narratives that can influence human perceptions and behaviour (Bingle 2023).

Yet there is a common language or matrix of key terms and concepts that has been defined by scholarship and refined by humanitarian practitioners (a synthesis has been provided in Box 2). This matrix is important for applying policy and international legal frameworks, yet it also reflects the complexity for practitioners on the ground to recognize and label different types of operations that manipulate information at a time of conflict (International Committee of the Red Cross [ICRC] 2021).

For the purpose of this paper, which focuses on situations of conflict, "information warfare" is defined as the collection, dissemination, manipulation, corruption and degradation of information with the goal of achieving strategic advantage over a conflict party and/or its population (Marlatt 2008; Prier 2017). Another comprehensive and more recent conceptualization of information warfare is "a struggle to control or deny the confidentiality, integrity, and availability of information in all its forms, ranging from raw data to complex concepts and ideas" (Bingle 2023, 6). As Morgan Bingle explains, "Offensively, information warfare occurs when one side within a conflict seeks to impose their desired information state on their adversary's information and affect how target individuals or populations interpret or learn from the information they possess or are collecting" (ibid.). Bingle stipulates that "the offensive actor can target at either the information itself or the individuals and larger group forming their target audience" (ibid.).

In this paper, "information" or "influence operations" are defined as "the strategic and calculated use of information and information-sharing systems to influence, disrupt, or divide society," for instance by involving "the collection of intelligence on specific targets, disinformation and propaganda campaigns, or the recruitment of online influencers" (Spink 2023, 48). Psychological warfare can be framed as "the planned use of propaganda and other psychological operations to influence the opinions, emotions, attitudes, and behavior of opposition groups" (ibid.). A useful definition of "adversarial information operations" is the one proposed by the "Oxford Statement on International Law Protections in Cyberspace: The Regulation of Information Operations and Activities" as "any coordinated or individual deployment of digital resources for cognitive purposes to change or reinforce attitudes or behaviours of the targeted audience."[3]

3 See www.elac.ox.ac.uk/the-oxford-process/the-statements-overview/the-oxford-statement-on-the-regulation-of-information-operations-and-activities/.

Box 2: Glossary

Definitions in this textbox are adapted from those of the ICRC, FP Analytics and Microsoft's Digital Front Lines report (2023) and the RAND Corporation, a non-profit global policy think tank that provides research and analysis to help improve policy and decision making. For terms related to technologies and their military uses, the author referred to research provided by Microsoft and RAND. The author used ICRC's insights for terms related to conflict, particularly the 2021 ICRC report *Harmful Information — Misinformation, Disinformation and Hate Speech in Armed Conflict and Other Situations of Violence* (ICRC 2021).

Misinformation: False information that is spread by individuals who believe the information to be true or who have not taken the time to verify it.

Disinformation: False information that is fabricated or disseminated with malicious intent. This can include terms such as propaganda and information operations.

Propaganda: Propaganda refers to information, often inaccurate or misleading, that is used to promote a specific viewpoint or influence a target audience. It might include elements of truth but presents them in a biased way to undermine the credibility or reputation of an opponent. When digital advertising, social media algorithms or other exploitative tactics are employed to spread propaganda, it becomes known as computational propaganda. This form of propaganda can also be used to target, recruit, radicalize and coordinate activities among potential supporters of extremist ideologies, a process commonly referred to as online radicalization and recruitment.

Hate speech: All forms of expression, including text, images, audio or video, that incite, promote or justify hatred and violence based on intolerance toward identity traits such as gender, religion, ethnicity or sexual orientation. This speech often blends misinformation, disinformation and rumours, and is manipulated by its perpetrators to fuel animosity. Utilizing both traditional and digital communication channels, hate speech exacerbates tensions between groups and can incite violence against individuals based on their identity.

Dual-use technologies: Technologies that have a primary civilian and commercial application, but also have the potential to be weaponized or used for military applications.

AI: The simulation of human intelligence in machines that are programmed to think and learn like humans. These machines can perform tasks that typically require human cognitive functions, such as visual perception, speech recognition, decision making and language translation.

Generative AI: AI systems that use advanced algorithms, such as generative adversarial networks and transformer models, to create new, realistic content, such as text, images and audio, that is often indistinguishable from human-generated content.

Foundational models: Large-scale AI models trained on broad data sets that can be fine-tuned for a variety of specific tasks. These models serve as a base or "foundation" upon which specialized models for specific applications can be built. The term has become prominent with the rise of large language models (LLMs) such as GPT-4, which are pre-trained on vast amounts of text data and can be adapted for tasks ranging from language translation to sentiment analysis with relatively little additional training.

Deepfake: An image or recording that has been convincingly altered and manipulated to misrepresent a person as doing or saying something that was not actually done or said.

Grey-zone tactics: The acts of state parties in relation to a dispute that maintain high-level diplomatic relations while interacting antagonistically below the threshold of war.

Hybrid warfare: The use of non-military tactics alongside conventional kinetic warfare to achieve foreign policy goals.

Technical Section

The technical section will analyze the potential of AI technologies, in particular generative AI, to influence and shape human behaviour at both the individual and population levels in situations of armed conflict and advanced geopolitical confrontations. This section aims to answer the following questions: What are the core converging AI capabilities and what is the critical leap achieved by generative AI? How can the confluence of these AI techniques act as a catalyst to amplify information warfare? How can these emerging trends in AI-led information warfare be harnessed by threat actors in conflict situations? And what are the potential impacts and reverberating effects on civilian populations, as well as on military forces and combatant strategies?

Core Converging AI Capabilities: What Are Converging AI Capabilities and What Is the Critical Leap Achieved by Generative AI?

The current AI revolution builds on a confluence of techniques and capabilities. Foundational AI models are systems that can learn to optimize large-scale data analysis, identify and classify patterns, structures and anomalies in vast data troves, and turn those insights into representations and predictions (Moy and Gradon 2023; Feldstein 2023). They are called "foundational" models because they serve as the groundwork for a wide range of AI applications, providing general purpose representations of data that can be fine-tuned or extended for specific tasks. Combined with an array of data-capture and sensing technologies, these models can be used to analyze a broad range of features and variations in a heterogeneity of data sets, from general image and text/language, to more precise features such as biometrics, human emotions and actions (Pauwels 2020b).

Generative AI models leverage the representations and features learned by foundational models to generate new content (text, images, narratives, videos and even music) that exhibits similar characteristics and patterns as the training data. For example, a generative AI model trained on images of human faces can generate new, photorealistic faces that closely resemble those in the training data set. LLMs are specifically designed to generate human-like text by analyzing vast amounts of language data.

Generative AI models have the potential to become increasingly autonomous, functioning as AI personal assistants and learning from different domains of human experience and expertise. For instance, LLMs have demonstrated the capacity to support laboratory work by providing options for building biological design and outsourcing complex tasks to adequate bio-foundries (Sandbrink 2023; Carter et al. 2023). In cybersecurity, generative AI models can learn from vast amounts of historical data on cyber incidents and predict future threats (Stanham 2023). Generative AI models also power AI decision-support systems that optimize data analysis and provide recommendations and predictions to aid decision making in war (Stewart and Hinds 2023).

Generative AI models are not only converging with dual-use expertise and other emerging technologies, but are also merging with our daily experiences, monitoring how humans live, move and feel. By progressively learning and simulating human inputs and behaviours, generative AI systems promise to develop dynamic content and sustain interactions that imitate the features of human conversations and, to some extent, relationships (Feldstein 2023; Hiebert 2024). With generative AI, the critical leap forward will likely come from both its increased autonomy and new capacity to capture, simulate and interact with human behaviours (Marcellino et al. 2023). These trends may amplify dual-use potential and unpredictability, with resulting consequences that are difficult to anticipate, mitigate and control.

Box 3: Information Operations and Their Harmful Impact on Specific Groups

The interplay between armed conflict and disinformation is intricately intertwined with existing grievances, amplifying human suffering, stoking hatred and disproportionately impacting vulnerable groups (United Nations General Assembly 2022).

Studies have shown the disproportionate impact of disinformation on women, children, and lesbian, gay, bisexual, transgender and questioning persons. For instance, women and children can suffer both psychological and physical harm from being targets of misinformation, disinformation and hate speech (United Nations Human Rights Commission 2018a, 2018b). Information operations can contribute to physical harm, including sexual violence — for example, when hate speech incites violent attacks against children of a minority group (Ridout et al. 2019). It can also lead to psychological and social harm through online harassment and sexual abuse as well as through digital hate speech and geo-targeted threats (when hate speech includes exact information about where women, children and sexual and gender minority populations live) (ICRC 2021, 9). Threat actors may resort to online information manipulation in order to target women and children, who are isolated from their families and in need of humanitarian help, and lure them to specific locations for trafficking.

Minorities and marginalized racial and ethnic groups often bear the brunt of the destructive effects of information warfare. For instance, in conflicts in Myanmar (United Nations Human Rights Commission 2018a, 2018b) and Ethiopia (Jackson, Kassa and Townsend 2020), combatants have exploited mass communication platforms to incite hatred, dehumanize opponents and trigger violations of human rights. In past conflicts in Kenya, Nigeria and South Africa, political leaders have employed divisive and inflammatory rhetoric to deny established facts, escalate tensions and incriminate national, ethnic and religious groups (Pauwels 2020a). Refugees, internally displaced persons and migrants are frequently depicted as threats to national security or social cohesion, fuelling hatred against them.

Generative AI: A Revolution for Information Warfare? How Can the Confluence of AI Generative Techniques Act as a Catalyst to Amplify Information Warfare?

AI technologies, in particular generative AI models, are making information warfare both more powerful and more accessible. The capacity of generative AI, merged with diverse forms of human behavioural data capture, provides more impactful techniques to drastically improve, tailor scale up and even industrialize the offensive use of disinformation.

Behavioural profiling and influencing

We have entered a technological era where our private and collective experiences have become free material for behavioural surveillance (Zuboff 2019). Our "patterns of life" — our conversations and emotions, biometric features and behaviours — can now be turned into predictive insights to fuel information warfare. The vast amount of digital information now generated by populations means that more of these routine behaviours can be understood through AI computing. A confluence of AI functions and techniques makes it increasingly possible to analyze, classify, profile and, to some extent, predict and influence human behaviour (Pauwels 2020b). The global AI industry posits that significant amounts of raw information about human experience can be turned into actionable intelligence; in other words, a critical mass of behavioural insights allows for individuals to be influenced remotely. For instance, targeted advertisements and content can exploit psychological triggers to influence purchasing decisions, voting behaviours or social interactions, creating an environment where individual

autonomy is significantly compromised by externally engineered stimuli. In 2018, the revelation that Facebook and Meta platforms made the private data of about 87 million of its users available to the Trump campaign fuelled new levels of public anxiety about the ability of tech giants to exploit or monetize personal information (Raymond 2022).

By accessing a myriad of human insights within our digital networks, generative AI will learn to profile crowds, classify sentiment and preferences, and simulate expertise, emotions and authentic behaviours, thereby crafting content that can be tailored, personalized and evolved over time (Marcellino et al. 2023; Beauchamp-Mustafaga 2024; Feldstein 2023; Hiebert 2024). Today, an industry flourishes around digital personas that use forged speech and videos to impersonate individuals — including deceased ones — with the goal of furthering personal relationships (Yang 2024; Carballo 2023). Think of chat bots that harness people's social media footprints and biometrics in order to become life partners or online ghosts.

Engineered reality and authentic human-AI relationships

The technological leap brought by generative AI will increasingly blur the distinction between real and synthetic content and authentic interactions and impersonations, challenging both human perception and machine detection.

LLMs enable the creation of a large amount of unique, long-form, higher-quality deceptive messages that go beyond short texts to news stories and public discourses, marking an incremental improvement over previous methodology. Deepfake technology is another example, in which image and video generation relies on the convergence of various algorithmic architectures, including deep residual networks that can, with surprising accuracy, read human lips, synthesize speech and simulate facial expressions and bodily movements (Mubarak et al. 2023). Its capability includes altering facial features and expressions, gait and biometrics, as well as simulating behaviours on video in real time. If an individual has a digital footprint that includes, for example, talks and podcasts, deep residual networks are also able to reproduce a synthetic version of their voice. On the eve of an election, deepfake videos could falsely portray public officials being involved in criminal or unsavoury behaviours. For example, in October 2023, just two days before Slovakia's elections, a Facebook post featured an audio recording purportedly capturing a conversation between Michal Šimečka, leader of the liberal Progressive Slovakia party, and journalist Monika Tódová from the newspaper *Denník N* (Zuidijk 2023). The voices on the recording seemed to discuss plans to manipulate the election, including buying votes from the country's marginalized Roma community. The deepfake was intended to discredit a liberal candidate and bolster support for more conservative and populist factions. Despite quick interventions to expose the fabrication, the rapid circulation of the recording likely contributed to shifting public sentiment and shaping election outcomes in favour of populist forces.

Public panic could also be sowed by videos warning of non-existent epidemics, health safety scandals or widespread cyberattacks. In April 2023, the US Republican National Committee released a 30-second video featuring AI-generated images of President Joe Biden and Vice President Kamala Harris celebrating an election night victory (Dorn 2023). The video then depicted simulated scenes of chaos, including explosions in Taiwan, police in tactical gear patrolling San Francisco, an influx of migrants at the southern US border and deserted buildings on Wall Street. These forged incidents could potentially lead to international political or military escalations. With the proliferation of sophisticated deepfake videos, combined with deepfake backstories and cover-ups, even qualified news reporters, decision makers and diplomats will increasingly struggle to parse propaganda and disinformation from real news. Already, lawmakers across the globe are being targeted for their positioning over geostrategic competitions and conflicts. In 2024, *The New York Times* reported that an Israeli political consulting firm called STOIC received US$2 million from Israel's Ministry of Diaspora Affairs to influence democratic members of the US Congress to ensure their support for Israel, at a time when many of these members are questioning continued American military support to Israel amid rising civilian casualties and suffering in Gaza (Jingnan 2024).

However, what truly sets generative AI apart is the potential for vast bot networks to convincingly mimic spontaneous human behaviour. Such automated networks can generate text, images and soon, with all likelihood, video and audio, bolstering the credibility of the messenger and the persuasiveness of the interaction (Marcellino

et al. 2023; Feldstein 2023; Brandt 2023). Individual customization is promised as the next breakthrough, with AI assistants increasingly mimicking genuine interpersonal relationships and potentially replacing or competing with human social bonds (Hiebert 2024). Increasingly, generative AI models will learn to create personalized content in real time through individual interactions with chatbots, leveraging granular population and user data to craft tailored messages that resonate with specific personas. As Kyle Hiebert has eloquently written, this trend could result in "digital siloed forms of existence," nihilism in relation to objective truth, political apathy and "erosion of civic engagement and social capital" (ibid.).

As they have infiltrated into our routines and daily lives, generative AI models will develop an emerging capacity for decision making, which could be used in dynamic relationships to progressively influence and take control over both targeted and larger audiences. This capacity to influence public opinion with misleading simulations and mobilize large swaths of populations around aggressive narratives could have powerful long-term implications for maintaining peace and security.

In general, the deployment of generative AI forgery technology will drastically alter the relationship between evidence and truth across journalism, criminal justice, conflict investigations, political mediation and diplomacy. By eroding the sense of truth and trust between citizens and the state — and indeed among states — generative AI's misuse and abuse could become deeply corrosive to democracies, global expertise and international governance systems.

Industrialization and privatization of information warfare

Through sustained campaigns and relying on human-like interactions, generative AI can be used to automate content dissemination at low cost and on a large industrial scale.

Large private sector groups and governments are already ramping up investments to develop current and more refined generative AI systems. Leading tech nations will have unrivalled advantage as they already power global networks, such as Meta, Instagram and TikTok, and can exploit massive sources of behavioural surplus about populations and subgroups. Yet, increasingly, data troves, including routine and sensitive information about civilians, are monetized and acquired by private sector offensive actors, proxies and cybercriminal groups (Pauwels 2024). Previous examples of unregulated, irresponsible innovation have shown potential risks. For example, Clearview AI, the controversial facial recognition company, has developed a powerful facial recognition algorithm capable of identifying individuals from images taken from the internet. The company claims to have amassed a database of billions of images sourced from social media platforms, websites and other online sources (Hart 2022). The technology works by comparing facial images from these sources against those in its database to generate potential matches, along with links to the source images. This capability has raised serious concerns about privacy, civil liberties and human rights. Techniques such as algorithm and data-exploitation leading to misuse and abuse are afforded to both state and non-state actors.

Non-state actors and proxies have increasingly gained access to some generative AI capacities through decentralized Web3 platforms, as well as acquisition in other ungoverned markets. Similarly to trends in cybercrime and the cyber arms race, a number of generative AI systems are being customized, repurposed through open-source platforms and acquired within dark web communities and underground marketplaces (Pauwels 2024). Open-source AI research actually boomed in 2023, with AI-related GitHub increasing by nearly 60 percent compared to 2022 (Maslej et al. 2023). Meta has published its generative AI model (called "Llama 3") as an open source, which means that the model's source code can be modified and repurposed. Dubbed "WormGPT" and "FraudGPT," open-source versions of OpenAI's GPT model are monetized on the dark web and may already have been repurposed in cyberattacks and fraudulent hacks (Wirtschafter 2024).

Past research on the "industrialization" of cyber offence highlights what experts have detected on the ground: increased forms of trading, collaboration and outsourcing between threat actors, including state proxies, mercenaries and cybercrime groups. Similar dynamics could accelerate and amplify an already emerging trend: the industrialization and privatization of information warfare (McGuire 2021; Pauwels 2024).

The implications of rapidly expanding and unregulated markets for information warfare will be corrosive to international peace and

security, with a potential rise in information operations leveraged by mercenary and terrorist groups. The availability of targeted influencing or large-scale disinformation to anyone who can afford it is already transforming how contemporary conflicts are fought. Both state and non-state actors are drastically empowered through information warfare, but the power relationship between these parties becomes less asymmetrical, with an increased diffusion of power. As a result, the potential beneficiaries of the industrialization of information operations may include private mercenary groups, terrorist groups, transnational illicit networks and proxy forces involved in conflict.

Such diffusion of cyber power could rapidly reach increasing numbers of private sector offensive actors and private groups associated with mercenary activity (Agranovich 2023). For instance, the former Wagner Group was actively involved in spreading global disinformation campaigns and leveraging influence operations (Marten 2022). Harmful implications have already been seen through information operations waged in different countries across the globe. Terrorist organizations may acquire, exploit or outsource services to support their offensive agendas, resulting in an increasing correlation between criminal accessibility and mercenary and terrorist capability. Violent extremist groups and criminal organizations, from Hamas and Boko Haram to Mexican drug cartels, have relied on cyberespionage to infiltrate governments and collect private information about intelligence personnel (Wirtschafter 2024).

Waging Information Warfare: How Can These Emerging Trends in AI-Led Information Warfare Be Harnessed by Threat Actors in Conflict Situations? And What Are the Potential Impacts on Civilian Populations and Military Forces and Their Combatant Strategies?

Information operations have emerged as a powerful threat with the goal to undermine civilian resilience of populations in conflict situations — even in combination with kinetic attacks — and to manipulate public opinion within and beyond national borders. The development of generative AI models means that information operations will likely become more adaptive, interactive and manipulative, waged with both precision and iteration at personal, local and global scales. While these attacks can be aimed at both military forces and civilian populations, a recent body of research shows that civilians are increasingly targeted by hostile operations that manipulate information critical for their survival, and influencing their behaviours to the point of causing harm and undermining their security and well-being (Katz 2021; Morris 2024; Burt 2023; Khan 2023; Lahmann 2020; Feldstein 2023; United Nations General Assembly 2022; see Box 3).

Targeting Primarily Civilian Populations

Psychological Operations Undermining Civilian Security in Conflict

In conflict zones, access to reliable information plays a crucial role in civilian protection. In the words of Irene Khan, the UN Special Rapporteur on the promotion and protection of freedom of opinion and expression, "the freedom of opinion and expression, including the right to seek, receive and disseminate diverse sources of information must be upheld by States in times of crises and armed conflict as a precious 'survival right' on which people's lives, health, safety, security and dignity depend" (United Nations General Assembly 2022, art. 19(2), paras. 1 and 5). Individuals affected by conflict are particularly susceptible to the harmful effects of disinformation (ibid.) due to their dire living conditions, pervasive confrontations with violence and limited access to reliable information. Conflict zones have become powerful incubators for disinformation, and hostile state and non-state actors have effectively used disinformation to shape the narratives behind conflicts and impact civilians. Intense social fragmentation and weakening of public institutions in these settings amplifies the impact of disinformation, creating "digital siloed forms of existence" that reinforce the rapid and endemic tactics of information warfare (Hiebert 2024).

In this context, a troubling use of information and psychological warfare adopted by conflict parties involves the calculated manipulation of information critical to meeting human security needs, with

the goal of influencing the behaviours, emotional states and well-being of civilians. Local-level information and psychological operations have had some of the most harmful consequences for civilians — obscuring frontline developments, sowing panic, preventing or hindering civilian efforts to evacuate from conflict-affected areas and deceiving civilians about the availability and functioning of critical infrastructures and emergency services (Spink 2023). Combined with predictive behavioural and emotional analysis, the use of generative AI models could transform such operations into a pervasive and persuasive form of psychological warfare. We have witnessed in recent conflict situations cyber operations that aim to destroy or manipulate the integrity of strategic data sets and the industrial control systems related to cities' infrastructures (Pauwels 2024). Such destructive cyber capacity could be used in combination with targeted forms of information and psychological operations to undermine citizens' security and trust in needed critical systems.

These operations can have long-term implications for the mental health of civilians by inciting terror, high levels of anxiety and other distressing emotions or mental states. Targeted psychological operations can lead to paranoia, conspiratorial thinking, the constant apprehension that basic human security and family needs are not being met and a pervasive anxiety about death or injury (Katz 2021). These psychological harms, though harder to document, can induce long-term trauma. By engineering "trust disorders" with public institutions or humanitarian organizations, hostile actors also aim to incite dissent, destabilize society, exacerbate situations of emergency, and undermine the reputation of enemy institutions, including those serving civilians.

For example, in the war of aggression against Ukraine, Russian-affiliated actors have conducted hyper-local disinformation campaigns that have merged with military offensives, targeting specific geographic areas with precision (Giles 2023). In the lead up to and aftermath of Russia's full-scale invasion, interviews with experts conducted by the Center for Civilians in Conflict revealed that networks of Telegram profiles and channels emerged at the community and oblast levels and that pro-Russian operatives infiltrated local Viber groups, spreading disinformation tailored to these communities (Spink 2023). This hyper-local approach poses significant challenges for Ukrainian officials and civil society to monitor and counteract. Unlike broader strategic narratives, which can be more easily detected, localized disinformation blends seamlessly with the chaotic flow of information during active conflict. This makes it difficult to discern deliberate disinformation from the misinformation that naturally arises in such volatile environments. The synchronization of these localized information operations with kinetic military actions has significantly amplified their impact (Burt 2023; Fedorov 2023). This strategy heightens confusion and panic among civilians precisely when they need to make rapid life-or-death decisions, exacerbating the already dire circumstances of those caught in the conflict.

In the first weeks of Russia's invasion of Ukraine, Russian actors launched a wave of disinformation about frontline developments, manipulating insights about areas under Russian control, including troop numbers and their movements (Spink 2023). Civilians were frequently misled with claims that local authorities had abandoned their roles or capitulated. Furthermore, they propagated alarming but fabricated reports of impending offensives, including fictitious threats of nuclear attacks and strikes on nuclear power plants, meant to induce widespread public fear and terror (ibid.).

Russian and Russian-affiliated actors have heavily focused their information operations on manipulating population movements, with the aim of influencing Ukrainian civilians to either stay in Russian-occupied areas or flee toward regions under Russian control (ibid.). These efforts often contribute to forced displacement, a violation of international humanitarian law (IHL) that encompasses coercion, fear and psychological pressure.[4] Central to these operations has been manipulating information about options for protection and roads for evacuation. By sowing doubt, these actors seek to hinder civilian attempts to escape, effectively trapping them in areas under Russian influence and exacerbating the humanitarian crisis. To a lesser extent, Russian disinformation efforts also aimed at undermining access to or the delivery of life-saving services by claiming that hospitals were overwhelmed and that food and electricity would be unavailable in certain zones. Finally, other targeted operations, often using video propaganda, were aimed at persuading Ukrainian parents in occupied

4 See https://ihl-databases.icrc.org/en/customary-ihl/v1/rule129.

territories to send their children to Russia, as part of a broader strategy to relocate as many Ukrainian children as possible to Russia, underscoring the civilian impact of Russia's information campaign in the conflict zone (Spink 2023).

The conflict between Israel and Hamas is another vivid illustration of the harmful impact on civilians that can come from tech-led and social media-driven disinformation, with platforms such as TikTok, X (formerly Twitter) and Instagram being inundated with AI-generated posts (Morris 2024). Social media users have unwittingly dispersed misinformation to multiple platforms, to the extent that even journalists have reported on the conflict basing their sources on manipulated information. Supporters of both Israel and Hamas have accused one another of victimizing vulnerable civilians by spreading forged images that picture the dead bodies of babies and children in order to produce emotional reactions (Klepper 2023). In some instances, pictures from previous conflicts or emergency disaster situations have been modified and presented as being current; in others, generative AI programs have synthesized images from scratch, including one of a baby crying amid bombing wreckage that went viral in the conflict's earliest days.

The rationale behind sophisticated and large-scale disinformation architecture is to immerse citizens in a virtual siloed reality in which they themselves become the producers of information and emotional manipulation. Interestingly, this tactic muddies who is supposed to carry the burden of intent behind waging information warfare.

In conflict situations, the consequences are corrosive as pervasive forms of disinformation undermine the reliability of all available information, creating chaos and hindering civilians' ability to make safe decisions (Morris 2024). For instance, disinformation campaigns about areas of Hamas operations misled civilians about safe areas, while similar tactics targeting the availability of essential supplies disrupted long-term survival planning. Tamer Morris explains that "when the Israeli government advised civilians to flee certain areas, no other information was provided to ensure safe evacuation, for example safe areas and corridors, implementing an atmosphere of chaos…this particular situation was further exacerbated as Israeli forces cut all telecommunications preventing civilians from sharing immediate information regarding safe passages or shelters, incapacitating their ability to make decisions" (ibid.). Humanitarian efforts, including the UN Relief and Works Agency and Médecins Sans Frontières, have also become the targets of disinformation, eroding public trust, complicating aid delivery and endangering workers.

Behavioural Control in Repressive Regimes and Intrastate Violence

Authoritarian states — sometimes in concert with private sector actors in the global security industry — may misuse and abuse AI and civilian data sets for social surveillance and control and ethnic and racial profiling, with the ultimate goal of amplifying propaganda efforts and manipulate populations. An increasing number of countries, including repressive regimes, are relying on AI and population data sets to monitor behaviours, implement social control and strengthen their surveillance apparatus (Feldstein 2023). There is a growing "tech assemblage" or "internet of bodies and minds" that can be harnessed to capture people's "behavioural surplus," involving internet and communication monitoring, mobile device hacking, computer interference, financial and geo-tracking, facial recognition, mobile biometric devices and "below-the-skin" technologies such as DNA sampling (Pauwels 2020b).

Adding generative AI to these converging data-capture techniques would allow for the direct targeting of population subgroups — such as partisans, dissenters, youth, women, ethnic majorities or minorities — with tailored, persuasive and interactive forms of propaganda and even psychological engineering. The goals of such targeting may be to inflame existing tensions and incite violence between ethnic and socioeconomic groups; track, deceive and silent opponents; recruit youth into information warfare and armed forces; subdue and repress human rights and women rights' efforts; and anticipate and manipulate sub-populations' movements during protests or social unrest.

Illiberal regimes and their proxies, as well as other violent actors, may combine behavioural monitoring with generative AI models to enhance the persuasive power and credibility of psychological operations that could use realistic impersonations and interactions to deceive specific population subgroups, including dissenters and human rights defenders. As witnessed in cybercrime, the combination of psychometric

tools and emotional engineering using personal data sets can help craft attacks so subtle that they are hardly recognizable as such (Pauwels 2020b).

Behavioural surveillance through algorithmic and cyber techniques is already a pervasive reality of intrastate violence and contemporary conflicts. These practices are reminiscent of the Syrian conflict that involved several cyber proxy groups, most prominently the Syrian Electronic Army (SEA) that acted in support of the government and President Bashar al-Assad. A 2021 report by cybersecurity and legal experts exposes how, "in conjunction with actively monitoring their own citizens, the Syrian regime, together with third party groups, is hacking websites and individuals critical of the regime" (The UIC John Marshall Law School International Human Rights Clinic [IHRC] and Access Now 2021, 1). The report continues, "Through 'phishing' operations, social engineering, malware downloads, and gaining access to passwords and networks through security force intimidation, the SEA and the Assad regime have used these practices to monitor and track down activists and human rights defenders in Syria, who are then tortured and killed" (ibid.).

In 2013, SEA members extracted from a standard messaging application the personal information (phone numbers, email addresses and contact details) of millions of people and leaked the data sets to the Syrian government (Kastrenakes 2013). Other attacks targeting social media platforms and messaging applications led to further breaches of civilians' sensitive data, including people's birthdays, personal serial numbers, ID cards, CVs and blood types. The report by IHRC and Access Now claimed that "the monitoring and hacking of devices are suspected to have informed kinetic operations that have cost the lives of many and undermined the crucial work being done by doctors and human rights defenders" (2021, 21). The deceptive tactics used by the SEA include social engineering and impersonation to manipulate anti-Assad activists into revealing the identities of dissidents and meeting locations. In the wake of such pervasive surveillance practices, surgeons and doctors have been advised not to provide medical mentorship over the internet to colleagues in Syria for fear of revealing the location of sheltered and underground hospitals (Baraniuk 2018).

Scaled-Up Information and Influence Operations in Ethnic Conflict

Hostile states and their proxies, violent extremists and other threat actors may increasingly rely on influence operations to increase political polarization and sow social unrest and ethnic conflict. Combined with predictive behavioural monitoring, generative AI models could identify the emotional triggers that push subgroups to violence and tailor disinformation campaigns and psychological manipulation techniques to be harnessed by factions in conflict, from ruling elites and political parties to terrorist groups. Lack of safeguards in social media networks and immersive digital spaces could enable state and non-state actors alike to manipulate individuals' deepest fears, hatreds and prejudices. For instance, violent extremist groups may spread false claims of violence committed by their enemies to inflame tensions and gain sympathy for their cause. When the Islamic State (IS) increased its power and visibility through social media, its violent propaganda, which used doctored videos and AI bots to magnify messaging, resulted in a wave of online emotional warfare (Ward 2018; Alfifi, Kaghazgaran and Caverlee 2018). The violent anti-Islamic backlash that followed was then instrumentalized for the group's recruiting strategies.

Applying generative AI to population behavioural data could drastically enhance methods and techniques in influence operations by supercharging the strategic communication environments in which conflicts play out. Both state and non-state actors can already feed their own narratives and mis- and disinformation to their constituents both within and across borders. Russian troll factories outsourced business to trolls in Ghana and Nigeria working to foment racial tensions around police brutality in the United States ahead of the 2020 election (Ward et al. 2020). In India's West Bengal region, Rohingya refugees have been demonized by the same kinds of extreme threats and online hate mongering that caused them to flee Myanmar (Goel and Rahman 2019). In Kenya and South Africa, disinformation and hate speech, manufactured in part by political elites, inflamed the racial and socioeconomic divisions that have plagued both countries for decades (Segal 2018). With AI technologies that can synthesize media from scratch, including graphically violent video propaganda, the art of emotional manipulation could become ever more powerful

and has the potential to inflict harm on specific ethnic groups and other vulnerable communities.

For example, since February 2022, pro-Russian social media outlets have propagated narratives aimed at inflaming social and linguistic tensions between population subgroups living in the western and eastern parts of Ukraine (Spink 2023). Information operations have raised concerns surrounding Russian-speaking internally displaced persons, alleging attacks, exorbitant rental fees and challenges in accessing education in western Ukraine. Additionally, stories have emerged claiming a disproportionate conscription of individuals from eastern regions of Ukraine into the military and unfair electricity rationing between western and eastern areas of the country. These narratives can significantly impact civilian well-being and social cohesion, perpetuating or worsening societal divisions, discrimination and violence.

Weakening Global Alliances and Public Support to Conflict Parties

In present and future information warfare, authoritarian and hostile states have a strategic interest in influencing large public audiences and distorting global perceptions of a conflict. Their goals include degrading access to trustworthy information, manipulating narratives, persuading global audiences of the futility of the fight and weakening strategic political alliances and public support for a conflict party. For instance, influence operations backed by Russia's Federal Security Service and other state-affiliated proxies have relied on "flooding social media platforms with misleading messages around the need for the 'denazification' of Ukraine and accusing the United States of creating bioweapons in clandestine laboratories in Ukraine" (Burt 2023, 15). As Annie Fixler underlines, "Russia has adjusted video evidence to deny war crimes, deployed operators on social media to create fake personas and news sites, and hacked user accounts to promulgate disinformation" (Fixler 2023, 11).

Beyond supercharging these global battles of influence, the development of generative AI models could accelerate and amplify synthetic data and forgeries to obfuscate criminal and state responsibility in the conduct of kinetic war and potential IHL violations. The industrialization of information warfare could result from generative AI's trends, including democratization, automation and outsourcing (with information warfare becoming a global and partially to fully automated "cybercrime as service"). Ultimately, all of these trends will make it increasingly complicated to trace the source and the supply chains of information operations, establish evidence and truth, and obtain material proof of instructions, directions or control in order to attribute state and/or criminal responsibility across jurisdictions.

Russia's strategic use of information operations to undermine Western support for Ukraine has been particularly evident in Germany, a key player in the European response to the conflict (Watts 2022). By exploiting societal divisions and economic fears, as well as leveraging a sophisticated blend of disinformation and propaganda via both traditional and social media, Russia has sought to erode the resolve of one of Ukraine's key European allies. Germany's significant Russian-speaking population, a legacy of historical migration patterns, has been the Russian media's key target, with tailored content that reinforces pro-Kremlin viewpoints and disseminates disinformation directly aligned with Moscow's strategic goals. This approach is intended not only to bolster support for Russia within this community, but also to create a potential internal pressure group that can influence broader public opinion and political discourse in Germany.

Targeting Military Personnel and Operations

Adversarial Information Manipulation to Thicken the Fog of War

In military domains, generative AI models are used to pioneer new ways to synthesize intelligence data and support human decision making, provide high-level strategic recommendations and new problem-solving techniques, generate different plans of attack and organize the jamming of enemy communications (Feldstein 2023; Stewart and Hinds 2023). Integrated into intelligence, surveillance and reconnaissance scenarios, generative AI models could support target tracking in drone missions. These AI capacities will also likely enhance the operation of a new wave of low-cost, adaptive and modular autonomous weapon systems that are designed to kill on a rapid scale (Federspiel et al. 2023). If misused

and abused in emerging forms of hybrid warfare that do not respect the rules of engagement by targeting non-military objectives, the number of resulting civilian harms could be unprecedented.

Military, legal and humanitarian experts have drawn attention to the promises of these kinds of AI decision support systems if used with a human-centred approach (Stewart and Hinds 2023; ICRC 2021): they could be used to foster the protection of civilians (by recognizing distinctive emblems and alerting forces about the presence of civilian populations), increase situational awareness and accelerate decision-making cycles. But these experts have also highlighted the potential problems and limitations of these generative AI systems, such as a greater reliance on rapid AI-generated analysis detached from battlefield observation and human experience, a subsequent accidental or intended escalation, the unpredictable risk-prone properties of these systems; and the challenges for humans interacting with AI reasoning (in particular, the problem of automation bias). All of these challenges show that for military forces, assessing and trusting the application of generative AI models will be a complicated decision.

In strategic military situations, one corrosive use of information warfare could be to wage adversarial attacks on generative AI models via the manipulation of data and signals. Such attacks could both poison the training data sets or the flow of insights captured into the system and manipulate its functioning, performance and predictive value. For example, ICRC experts posit "adversarial techniques [that] could conceivably be used in conflict to affect a targeting assistance system's source code such that it identifies school buses as enemy vehicles, with devastating consequences" (Stewart and Hinds 2023, 2). Adversarial attacks could virtually alter the performance and reliability of generative AI models in their different military configurations, from strategic and logistic planning and training and decision making, to intelligence, surveillance and reconnaissance, command and control, cyberoperations and autonomous weaponry. With adversarial attacks, the increasing dependence of military forces on generative AI could thicken the fog of war, undermining in-depth human understanding, situational awareness and compromising decision making, alternative options and on-the-ground operations.

What could ultimately be engineered for large-scale harm is the entire "intelligence life cycle" irrigating military operations, from inside tacit military expertise to large collected data sets, including extremely sensitive information about civilian populations and critical targets and infrastructure. Failure by military institutions to prevent adversarial attacks or the misuse of generative AI could be exploited subsequently by the enemy in further waves of information warfare targeted at destroying trust in local and global audiences.

The merging of the cyber arms and cybercrime industries is leading to a proliferation of dual-use expertise that can be harnessed to repurpose and re-engineer existing cybersecurity and AI systems. In this context, adversarial attacks could be performed by malicious actors with relatively sophisticated AI and cybersecurity skills or those able to acquire this knowledge, and could increasingly integrate the offensive tool kit of state and non-state violent actors, advanced persistent threats and cybercrime groups acting as proxies (Pauwels 2024).

Psychological Operations Undermining Resilience of Military Forces

The integration of generative AI into psychological operations offers unprecedented avenues to manipulate military forces in ways that can profoundly impact combat strategies, disrupt command and control and undermine resilience through emotional engineering. By simulating authentic human communication patterns and producing deepfake audio and video, AI can fabricate convincing messages purportedly from military leaders or trusted sources, potentially causing chaos and eroding resilience within opposing forces (Byman et al. 2023; Fecteau 2021).

AI-driven disinformation could be used to mislead enemy forces about strategic decisions, impacting combatant strategies. For instance, generative AI could create convincing but manipulated intelligence reports or communications that suggest a non-existent troop movement or supply route. By hacking a combination of personal and official communication channels and feeding this fabricated information to enemy analysts, military forces might be misdirected to either defend or attack the wrong locations, thereby compromising their operational effectiveness. Additionally, AI-generated deepfake videos or audio messages from supposed high-ranking

officers could order troops to execute defective strategies, leading to failures on the battlefield.

Generative AI can be instrumental in creating chaos within the command structure of an enemy force. By producing forged communications that appear to come from legitimate military sources, AI can sow confusion and mistrust among commanders and their subordinates (Fecteau 2021). For example, a deepfake video of a commanding officer issuing contradictory orders could lead to paralysis and indecision among troops. Furthermore, AI can generate false alerts about imminent threats, causing units to constantly reposition or retreat, thereby exhausting resources and morale. Automated bots and generative AI assistants could flood communication networks with manipulated reports, overwhelming the command's ability to process real-time information and making it difficult to execute coordinated manoeuvres (Lahmann 2020).

Generative AI can also tailor psychological operations to exploit specific vulnerabilities in an adversary's cultural or social fabric, undermining its resilience. By analyzing vast amounts of data from social media and other digital footprints, AI systems could identify key psychological triggers to craft personalized propaganda and emotional engineering strategies to manipulate soldiers. Generative AI could also create realistic but forged video or audio messages and social media posts from family members suggesting personal crises, health emergencies or threats at home. For example, a soldier might receive a highly convincing deepfake video call from a loved one, fabricated to appear as if they are in immediate danger, prompting distraction, distress and a compromised focus on their duties. Military units could also reveal their positions by attempting to connect with audiences at home. Additionally, AI-driven misinformation campaigns could spread rumours about widespread threats to soldiers' families, leading to heightened anxiety and decreased morale across the ranks.

By deploying AI and generative AI in these targeted and sophisticated ways, adversaries can significantly degrade the operational effectiveness, cohesion and resilience of military forces (Byman et al. 2023). These psychological operations exploit the very fabric of human trust and communication, making them a powerful tool in modern warfare.

Behavioural Engineering to Recruit Youth in Proxy Forces

Combining generative AI with the profiling of population data could lead to new forms of behavioural engineering that could become pathways to recruitment into cyber and information warfare, as well as into kinetic warfare waged by armed forces and non-state armed groups. Russia's ongoing war of aggression toward Ukraine confirms the proliferation of proxy groups that have engaged local and foreign remote hackers in offensive cyber and information operations on behalf of both parties to the conflict (Pauwels 2024). Cybersecurity experts have talked about the increase in young, cyber-skilled populations available for deployment by cyber proxy groups and states (McGuire 2021, 7). When adolescents are recruited or engaged in offensive cyber operations, their status may convert to that of an active combatant and they may become legitimate targets for retaliation. They may also unwittingly be part of conduct that involves war crimes.

The advent of generative AI, and its trends toward hyper-personalization and mentorship, could act as a catalyst to recruit different demographic groups, including youth, into information warfare, further blurring the lines between civilian and military functions and complicating responses by military forces (see Box 5). In the context of youth recruitment, socio-technical pressures could arise from the capacity for violent actors to exploit young users' data profiles and spheres of communication. The convergence of generative AI and cyber surveillance could amplify the ease with which these actors are able to reach out to vulnerable groups and scale both their digital involvement and physical enlistment with non-state armed actors. Through generative AI techniques in social networks and immersive digital spaces, young users could become psychologically isolated from traditional support systems and the victims of emotional targeting based on viral video propaganda, impersonations and group pressure on networking platforms.

Evidence from the field confirms that the phenomenon of online recruitment has continued to grow, contaminating rising tech platforms, such as TikTok, and reaching ever younger generations (Pandith and Ware 2021; Meisenzahl 2019). In recent years, governments and intelligence services have gathered evidence that gaming platforms that incorporate voice-to-voice video conferencing, chat and messaging services are incubators for non-

state actors and violent groups to communicate propaganda and groom young recruits across different regions and cultures (Concentric 2019). Exploiting online gaming platforms is a method reportedly used by a diversity of actors, including IS, the Lebanese group Hezbollah and white supremacist groups across Europe and the United States. Security firms have reported that mentorship on how to maximize gaming platforms to profile and enlist new members is part of strategic recruiting discussions on IS's deep web forum (ibid.). As Joseph Guay and his co-authors write, "Social media can also be a vehicle to facilitate both kinetic and digitally derived forms of violence in which cyber militias have engaged in online defamation campaigns and have weaponized rumours and false information to incite panic and/or violence" (Guay et al. 2019, 52).

Singer and Brooking have powerfully summarized how the weaponization of social media has "represented a momentous development in the history of conflict" (Singer and Brooking 2018, 9). Young internet users have been instrumentalized in online "Twitter wars" that could help shape their perceptions on real conflicts taking place on the ground. When IS increased its visibility and reach through social media, its violent propaganda, which sometimes features children perpetrating executions and other acts of violence, was further exploited to fundraise as well as recruit and train new members (Almohammad 2018). With youth constituting a specific demographic that has been increasingly targeted as future perpetrators of disinformation and hate speech, network effects and immersive digital spaces could then augment the risk of adolescents becoming increasingly active in recruiting others to participate in information warfare and physical acts of violence. In the near future, enhanced by mentorship and personalized interactions brought by generative AI, these recruitment dynamics would result in both individual violations of rights and collective harms, and would have potentially corrosive implications for military forces' plans.

Scenario: Information Warfare on Biological Threats

An even more powerful and radical shift will come from the convergence of generative AI with other dual-use technologies and its integration within infrastructures that are critical to national security. For instance, generative AI is used in cybersecurity to improve threat detection and response and predict future polymorphic attacks. But the same AI models could also help conceptualize and plan how to modify, deliver and disseminate biological agents (Sandbrink 2023). In the near future, it is also likely that generative AI models will merge different types of "live scientific mentorship" (text, audio, immersive video) that could increasingly support lab work for less sophisticated threat actors.

An increasing number of threat actors could therefore access and process dual-use knowledge that could subsequently be used in information warfare to credibly impact both military and civilian populations. What is at stake is the weaponization of dual-use knowledge itself, and possibly all forms of dual-use expertise developed by human civilization. This is particularly salient in the case of AI and biotechnologies, for which there is a false dichotomy of dual use, as almost all aspects of both of these technologies that are deployed in service of human security can also be subverted for misuse by hostile actors.

The convergence of generative AI and biotechnology could be weaponized to create disinformation campaigns about biological threats, tailored to destabilize civilian populations and undermine trust in public health systems and governing institutions (Gisselsson 2022). With these aims in mind, information warfare would not be waged to impact military contingents and achieve kinetic advantage, but instead to cause a massive psychological impact on civilian populations and weaken allied countries' support. The below hypothetical scenario explores how such tactics might reach strategic effects in a situation of protracted conflict between two states (see Box 4).

Imagine a conflict in which a hostile state leverages generative AI to craft a sophisticated disinformation campaign, exploiting advances in biotechnology.

The adversary aims to create panic and chaos within the civilian population of a targeted nation by fabricating threats of biological attacks.

Phase 1 — Crafting the narrative: Relying on the expertise and mentorship of a generative AI lab assistant, the hostile actors access critical, but vulgarized knowledge about biotechnology and learn enough about how to conceptualize and plan a simulated release of biological agents. Then, using LLMs, the hostile actors create detailed and realistic scenarios involving the release of a genetically engineered pathogen. These scenarios are carefully tailored to resonate with existing fears and vulnerabilities within the target population. Generative AI models synthesize a series of fabricated news articles, social media posts and manipulated scientific reports, all suggesting that a lethal, highly contagious virus has been released in key urban centres. For example, AI-generated content might describe a supposed outbreak of this novel virus in a major city, combined with deepfake videos quoting medical experts and government officials, showing overwhelmed hospitals and quarantine zones. The content would then be distributed through a variety of channels, including social media platforms, fake news websites and even hacked legitimate news outlets.

Phase 2 — Amplifying the disinformation: To ensure that the disinformation spreads rapidly, the adversary employs bots and trolls to share and comment on the AI-generated content. These automated operatives flood social media with alarming posts, creating trending topics and hashtags that draw widespread attention. Enhanced by generative AI, the bots engage in personal and group discussions using scientific and vulgarized arguments, counteracting attempts to stop disinformation with formal investigation and posing as concerned citizens or medical/expert whistle-blowers, sometimes even impersonating individuals' close contacts: all tactics designed to amplify a sense of urgency and panic. In tandem, the adversary hacks into local news stations and inserts forged breaking news segments about the outbreak. These segments are crafted to look as authentic as possible, with realistic graphics and credible-sounding reports, further blurring the line between reality and fiction.

Phase 3 — Exploiting AI and biotechnology: To lend credence to their claims, the hostile state uses a generative AI lab assistant and access AI-led bio-design tools to create real but non-lethal and unfamiliar strains of biological agents. These agents are released in select locations, causing noticeable symptoms similar to those described in the disinformation campaign. When people in these areas start experiencing symptoms, it fuels the belief that the fabricated outbreak is real. Moreover, the adversary plants and releases (using micro-drone technologies) manipulated biological samples in hospitals and research labs, showing the presence of the engineered pathogen. These samples are designed to be detected by standard testing methods, leading to false positives that further corroborate the forged reports. Cyber proxies of the hostile state conduct a series of sophisticated adversarial attacks on health and biotech infrastructures to disrupt the production of prophylactic medicines, as well as to suppress or manipulate the content of data sets used in medical and epidemiological reporting.

Phase 4 — Medical and psychological impact on civilians: As news of the outbreak spreads, public trust in health authorities begins to erode. People flock to hospitals, overwhelming health-care systems that are already strained from the protracted conflict. Pharmacies run out of basic medical supplies as panicked citizens try to stockpile what remains of medications and personal protective equipment. strained from the protracted conflict. Pharmacies run out of basic medical supplies as panicked citizens try to stockpile what remains of medications and personal protective equipment.

Box 4: Reverberating Impacts on Vulnerable Groups

For groups in situations of vulnerability, including children, pregnant women and persons with disabilities, access to specific medical, child and disability care services could be strictly limited and the subsequent health and psychological impact harmful. For instance, in contexts where disinformation would blur the nature and scale of a biological attack, pregnant women and children may not benefit from appropriate medical countermeasures or may undergo unnecessary stressful procedures. For children and women in situations of vulnerability, the reverberating implications of a public health crisis — caused by information warfare around a biological event — may also include hindering access to food support services, schools and other needed resources, exacerbating the harmful psychological impact.

The disinformation campaign also targets specific communities with tailored messages, exploiting existing socioeconomic and ethnic tensions. For instance, messages in predominantly immigrant neighbourhoods might suggest that the outbreak is being used as a pretext to enforce repressive controls and commit violence. In rural areas, the disinformation might claim that urban elites are being prioritized for treatment and vaccines. The resulting chaos hampers the government's ability to respond effectively. Emergency services are stretched thin, and misinformation about safe practices spreads rapidly, undermining public health efforts. Civil unrest begins to simmer as people demand answers and accountability from their leaders. Public health institutions struggle to regain their authority. General vaccination rates drop as anti-vaccine sentiments, bolstered by the disinformation campaign, take root. The societal divisions exacerbated by the tailored messages deepen, making it harder for the nation to recover and heal. In allied countries of the victim state, levels of political and public support drastically drop as paralysis is fuelled by fears of a spreading epidemic amid a lack of understanding on the origin of the biological threat.

These converging security risks would have corrosive implications for every country, but particularly those that have poor and outdated medical, biotech and cyber infrastructure or those that have a limited capacity to protect their vulnerable populations from the weaponization of pandemic and technological threats in situations of protracted conflict. The COVID-19 pandemic has provided state and non-state hostile actors with a real-time window into societies' strengths and weaknesses in emergency situations. The pandemic has shown how a biological threat could break down hospitals and food supply chains, shatter citizens' trust in public institutions and bring social unrest, disinformation and even violence. In a similar way, information operations leveraging access to dual-use expertise and mentorship provided by generative AI and threatening large-scale casualties could be used to multiply the threat in hybrid warfare scenarios. The most enduring harm would be to civilian resilience and trust: trust in public health institutions, emergency data systems, laboratories, hospitals and critical infrastructures. As generative AI learns to democratize strategic military and civilian expertise and tacit knowledge in complex technological domains, such capacity will change not only the scale, but also the nature and power of information warfare, impacting dual-use knowledge asymmetries between threat actors involved in conflicts.

Legal Section and Concluding Thoughts

Information warfare has become a powerful business and a pervasive threat, global in scope with real and unprecedented ramifications for the security and survival of civilian populations in armed conflict.

Three sobering observations can be made based on how information warfare is waged in modern conflicts. First, with Russia's extensive use of hybrid warfare techniques in the war in Ukraine, we witness how information and psychological operations can be combined with cyberattacks and integrated within kinetic warfare.

Second, to a more pervasive extent than ever before, the manipulation of information is purportedly designed to significantly impair civilians' decision-making process for self-protection and survival (Morris 2024; Katz 2021; United Nations General Assembly 2022). When information is weaponized that way, it impacts critical elements of civilian protection, causing direct and indirect harm to populations, in particular vulnerable groups such as women, youth/children as well as humanitarian and emergency personnel. As highlighted by Khan, modern armed conflicts and information and psychological operations, including those inciting violence, increasingly target civilian populations rather than military forces (United Nations General Assembly 2022). It follows from this deeply worrisome trend that the core doctrine regulating armed conflicts within IHL — the protection of the "person" — is disregarded and now often violated.

Third, the advent of generative AI will rapidly lead to an industrialization of information warfare, giving states and their proxies, non-state armed groups and other violent actors, enhanced, adaptive, self-refining techniques to influence and manipulate human behaviour in conflict. This new diffusion of power will first manifest through dynamic, interactive and persuasive ways to influence populations captured in digitally siloed forms of existence (Hiebert 2024). In future conflicts, there could be even more techniques enhanced by generative AI to manipulate people's ability for self-determination, self-protection and decision making in emergencies. Yet, as shown in the scenario, another wave of implications will come from the convergence of generative AI with other sensitive technologies and subsequent democratized access to dual-use knowledge and expertise, which could be exploited in future information warfare.

Increasingly, international legal experts recognize the need and urgency of clarifying the existing rules imposed by IHL on offensive information operations (Gisel, Rodenhäuser and Dormann 2020; FP Analytics 2023). It is equally urgent to weigh whether the current international legal framework adequately captures the humanitarian and civilian protection needs that arise from waging information warfare in conflict. These goals go beyond the purpose of this paper. Yet, it proposes to succinctly review some of the protective measures afforded in IHL and some of the legal ambiguities and critical protection gaps.

IHL

Manipulating information is not a new phenomenon of warfare. As long as they infringe no rule of international law applicable in armed conflict, deceptive information and psychological operations have been allowed in past hostilities, including misinformation as a ruse of war and the use of propaganda targeting civilians (Rodenhäuser and D'Cunha 2023; Katz 2021; Lahmann 2020). Yet, as eloquently underlined by Morris, "while there is no doubt that ruses of war and propaganda are permissible under IHL, this does not mean that all deceptive conduct is legal" and "this does not consequentially permit all information warfare in the future" (Morris 2024, 2). In the words of ICRC legal experts, "we must recall that there is a red line between an information operation that complies with IHL and one that violates it" (Rodenhäuser and D'Cunha 2023, 1). Since IHL is fundamentally aimed at protecting civilians, its provisions should be interpreted with this principle in mind. Therefore, information warfare must be governed by conflict parties' obligations to civilians under IHL: these obligations bind both state and non-state parties, including proxy actors, political parties or other forms of civilian leadership.

The prohibition to encourage unlawful violence entails that "civilian or military leadership of a party to an armed conflict must not order or encourage IHL violations by their own forces" or by groups of civilians when the commission of such violence is foreseeable (ibid.). To illustrate, a state party to conflict would violate its obligations under IHL if it conducted information and/or psychological operations to incite combatants or civilians to attack and harm other civilians and civilian objects

(Lahmann 2020). Information operations that incite or intently lead to violent attacks against humanitarian organizations are also prohibited under IHL. Conflict situations that involve inter-ethnic tensions, the proliferation of proxies and the tacit reliance on armed groups as surrogates for attacks could be prone to information warfare as incitement to violence.

A limited range of severely harmful types of information and psychological operations could be under the protective reach of IHL if it amounts to prohibited acts or threats of violence, the primary purpose of which is to spread terror among the civilian population. Two considerations may limit the effective application of this rule to modern information warfare (Lahmann 2020). First, information or psychological operations would not meet the threshold if they do not come with an actual or threatened act of violence, which might disqualify many operations even if they result in extreme fear among the civilian population. Second, legal experts would need to demonstrate that the main purpose of the act or threat of violence is to spread terror and that no other motives or objectives are predominant. The case study described in this paper presents a situation where the spreading of fear and terror related to biological threats is exploited for civilian destabilization and for weakening their trust in experts and governing institutions. In regard to the two above thresholds, it remains unclear whether such a scenario would clearly qualify as an act or threat of violence to terrorize the population, even if its impact could cause harm to civilians. While most instances of information warfare may not be so clearcut as in the following example, ICRC legal experts provide an illustration involving a cyber intrusion into digital networks that "propagate false air raid alarms" to "keep inhabitants in a state of terror, or to displace them" (Rodenhäuser and D'Cunha 2023, 3). When the target of spreading terror is military forces, IHL also provides certain limits, including that "threatening to kill, rape, torture, or ill-treat captured or wounded soldiers is a violation of IHL" (ibid.).

Legal ambiguities remain as to whether a certain level and type of information operations may meet the "attack" threshold under IHL, subjecting it to the rules on targeting, such as the principle of distinction, proportionality and precaution (Lahmann 2020). It is relevant to compare with existing discussions on what constitutes a cyberattack in IHL. Important technical questions persist about how to define and qualify — in the context of an armed conflict — technical terms such as "attack" when they rely exclusively on cyber and digital means. Yet, it is increasingly recognized that cyberoperations designed to bring physical destruction or death meet the attack threshold. In its 2020 position paper, the ICRC underlines the importance of considering "harm due to the foreseeable direct and indirect (or reverberating) effects of an attack, for example, the death of patients in intensive care-units caused by a cyberoperation on an electricity network that results in cutting off a hospital's electricity supply" (Gisel, Rodenhäuser and Dormann 2020, 313). Consensus among states is still lacking as to whether cyberoperations that would not cause physical damage but would result in disruption and loss of essential services, or in erosion of public trust in critical systems, would qualify as an attack and thus violate IHL. Such discussion is relevant to this paper's case study in which information operations are used in a conflict to manipulate knowledge and information about pretend biological threats to the extent of destabilizing civilians, leading them not to seek or trust medical care and pushing them to endanger their health.

While the issue is still debated, some experts argue that "just like other types of military violence, if the causal nexus between an instance of disinformation and physical harm is sufficiently strong so as to render such operation an attack, it must respect the distinction, precaution, and proportionality triad" (Lahmann 2020, 1241). It is likely that legal ambiguities would remain as to the "causal nexus," as well as to the matter of scale and effect and meeting thresholds.

Box 5: Recruiting Youth in Information Warfare

To prevent youth's recruitment and use in information warfare, the CRC Optional Protocol on the Involvement of Children in Armed Conflict, relevant Security Council Resolutions and other related normative standards (Paris Principles) could serve as a basis for legal interpretation and protection of children and adolescents. In addition, the International Criminal Court (ICC) includes in its list of war crimes the active involvement of children in hostilities. Further legal interpretation would be needed to determine under which specific conditions recruitment of adolescents and children into information operations would be prohibited under the Optional Protocol and other mechanisms. In particular, legal experts would need to clarify whether recruiting children to become perpetrators of offensive information operations could constitute, in certain circumstances, direct participation in ongoing hostilities. To ensure that children are not recruited or used in conflicts, including armed conflicts, through cyberspace, the Committee on the Rights of the Child encourages states parties to better control, even criminalize and sanction, the forms of behavioural targeting and grooming of children that are enabled by digital technologies on social networking platforms and online games (United Nations Convention on Rights of a Child 2021).

International Criminal Law

A limited category of harmful cyber and information operations in situations of armed conflict may also be regulated under international criminal law, which applies to any natural person who commits an international crime. Under this regime, individuals or groups engaging in information warfare may be prosecuted for conducting information operations that would constitute war crimes, crimes against humanity and genocide. To prove individual responsibility for these international crimes, two elements have to be established: *actus reus* (the physical parts of the crime) and *mens rea* (the intent to commit the crime). The principle of command responsibility (article 28 of the Rome Statute), established in customary international law, stipulates that military commanders may be held criminally responsible for crimes committed by armed forces under their effective command and control.[5]

For several reasons, discussions on how offensive cyberoperations could be regulated under the Rome Statute are relevant to information warfare. First, different definitions of information warfare coexist with several aspects of cyberoperations, including the manipulation of information, ranging from raw data and signals to complex concepts and ideas. Second, information and psychological operations might become increasingly combined with cyberattacks and even integrated within kinetic warfare. In practice, there is a permeability between offensive cyberspace tactics, where data exfiltration and monetization, cyber intrusion and espionage can support and merge with methods of information warfare. For instance, as shown in the technical section, a cyber intrusion would likely be needed to launch an adversarial attack that would compromise the functioning and the integrity of data within AI models in civilian or military systems. Under the impulse of convergence with generative AI, the permeability and overlap between cyber and information operations will likely increase. Third, in legal reasoning and proceedings, it might be more strategic to consider these "digital means" as having synergistic and cumulative impacts.

In 2019 and 2020, a Council of Advisers' Report on the Application of the Rome Statute of the ICC to Cyberwarfare provided critical insights into how the ICC may regulate cyber and information operations that have the potential to cause grave suffering for the civilian population, including suffering equal to that caused by the most serious international crimes (Permanent Mission of Liechtenstein to the United Nations 2021). For instance, members of the Council of Advisers confirmed that an adversarial attack altering or

5 Command responsibility is a jurisprudential doctrine in international criminal law permitting the prosecution of military commanders for war crimes perpetrated by their subordinates. The first legal implementations of command responsibility are found in the Hague Conventions IV and X. See ICC (2021, art. 28).

deleting civilian medical data may be considered a violation of IHL, and therefore possibly a war crime (ibid., 39). The council also specified conditions under which cyber and information operations could lead to crimes against humanity: for instance, by inflicting serious and systematic harm to the mental health of a targeted group to the extent that it would amount to torture or persecution (ibid., 65–67). In particular, the council agreed with the UN Special Rapporteur on torture and other cruel, inhuman or degrading treatment or punishment that "cybertechnology can also be used to inflict, or contribute to, severe mental suffering while avoiding the conduit of the physical body, most notably through intimidation, harassment, surveillance, public shaming and defamation, as well as appropriation, deletion or manipulation of information" (Bowcott 2020). Regarding the crime of genocide, members of the council concluded that cyber and information operations may not only contribute to severe psychological and mental harm, but also help initiate and amplify physical acts of violence that could threaten the destruction of a specific minority (Permanent Mission of Liechtenstein to the United Nations 2021, 80–88).

Critical Civilian Protection Gaps

What surfaces through legal analysis is that existing international legal frameworks might not be adequate and comprehensive enough to address the emerging issues posed by information warfare and the converging AI and dual-use technologies it involves. While a few provisions in existing IHL impose constraints on information and psychological operations, these rules rely on definitions, criteria and thresholds that do not necessarily reflect the way that information warfare is integrated with hybrid tactics in cyberspace and waged in modern armed conflict (Lahmann 2020). With a rigid approach to the application of IHL, we will face legal ambiguities and grey areas that allow information and psychological operations to continue harming civilians (Katz 2021; United Nations General Assembly 2022).

Several key challenges and critical civilian protection gaps persist that will require attention in the coming years. First, when information and psychological operations are conducted in armed conflict, IHL provisions might not adequately cover instances of harm that are pervasive but difficult to attribute and qualify legally, such as exposure to foreseeable violence, the manipulation of information undermining civilian security and well-being, and mental suffering. The conduct of information warfare can increasingly be automated and waged remotely via outsourcing to proxies, yet the rules of war apply to areas that are controlled by conflict parties (Katz 2021). Because disinformation often causes harm indirectly, it is unlikely to be classified as an attack or act of violence under IHL, nor would it be considered incitement unless it explicitly advocates violence or hostility. Next, although disinformation can cause direct harm to mental health, these injuries are difficult to assess and document in real time and are not adequately or sufficiently considered in IHL. Finally, while adversarial information operations can impede the function of civilian infrastructure, it often does so in ways that do not constitute an attack under existing legal definitions.

Second, existing international legal doctrines do not cover an array of offensive information and psychological operations because they do not clearly qualify as mere instances to terrorize or incite violence, even if they aim to significantly degrade the integrity of the information ecosystem during armed conflict (Lahmann 2020). The goals of these operations are to systematically target civilian populations by weakening resilience and trust in governing institutions, undermining self-determination and decision-making processes and even exploiting the democratization of dual-use expertise to wage powerful forms of information warfare that could lead to large-scale destabilization and insecurity. The nature, scope and impact of these manipulative operations, along with their enduring divisive and corrosive effects on public trust and societal stability, underscore the need for greater scrutiny and attention, including and particularly when they are waged during armed conflict.

And in the case of information warfare that targets critical elements of civilian protection and infrastructures and integrates with kinetic military operations, there might be an argument to make about the need to assess their intensity and potential cumulative impact on resilience and survival and the subsequent physical and mental suffering it has imposed on the civilian population. Similar legal discussions about considering cumulative quantitative and qualitative impacts of offensive cyberoperations in armed conflict — and their qualification as "cyber war crimes" — are happening under the auspices of the ICC (Khan 2023). The prosecutor of the ICC highlights that,

"as states and other actors increasingly resort to operations in cyberspace, this new and rapidly developing means of statecraft and warfare can be misused to carry out or facilitate war crimes, crimes against humanity, genocide, and even the aggression of a state against another" (ibid., 50). Yet, ICC proceedings require very high evidentiary standards for attribution, and this is particularly relevant to the involvement of proxies in information warfare. How to qualify, document and attribute international crimes in the digital context and how to proceed across jurisdictions will continue to create legal ambiguities and challenges. We need a whole-of-society response for these types of attacks that affect us all.

Third, with the advent of generative AI and other dual-use technologies, as well as the weaponization of cyber capabilities, we face the prospect of a rapid proliferation, commoditization and privatization of information warfare. Already, nation-states are outsourcing information and psychological operations to a growing number of cyber proxy actors, including in armed conflict. At the same time, cyber proxy activity is becoming increasingly difficult to decrypt, trace and attribute (Pauwels 2024). The frameworks used to categorize forms of deputization in cyberspace do not adequately capture the increased permeability and intense knowledge and technical transfer that exist among non-state actors, as well as between state and non-state actors (ibid.). The polymorphous, multi-jurisdictional nature of cyber proxy activity therefore drastically complicates technical and, to an even greater degree, legal attribution of wrongful conduct in cyberspace. The consequence is that deniability remains more than ever a winning strategy for states using cyber proxies to wage information warfare and advance their geostrategic interests, particularly in the absence of an independent and multilaterally recognized attribution authority. The normative gap that will persist for the coming years in this regard — coupled with the potential involvement of decentralized private actors in the design, management and procurement of generative AI and other dual-use technologies — will give allowance to new types of abuses being left unaddressed.

Fourth, complex accountability and compliance challenges should raise questions about the role of both military institutions and the defence and civilian private sector in strengthening responsible innovation and protecting governments, populations and industries. Private sector actors not only bear a major responsibility, but also are best placed to use oversight and foresight in the rapid development of generative AI, its convergence with other dual-use technologies and its integration within increasingly blurred military and civilian critical infrastructures. There is a need for the private sector to recognize its role in collaborating with military leadership and disarmament architectures to prepare for the diffusion of power in conflict that will come from the proliferation and democratization of AI and converging technologies.

Preliminary Recommendations

As AI and generative AI systems reshape how knowledge, expertise and information is used and potentially manipulated in conflict and the grey zone between war and peace, now is the time to think forward and assess risks, vulnerabilities and forms of resilience. While there will be specific implications for military forces and strategic thinking, prevention and resilience will depend on a whole-of-society response.

It is crucial to adopt a multi-stakeholder, collaborative strategy that includes the active participation of civil society, traditional media, governments and military forces, international entities and digital corporations. It is equally important that states dialogue with rights holders and civil society to forge a vision of how to best build social resilience against information manipulation. States will need to continuously map how these new deception tools influence public discourse and opinion. They will also need to foster cybersecurity and (bio)technological literacy in their civilian populations. As Khan emphasizes, "more attention should be given in fragile situations to media information and digital literacy, particularly for young people, women, the elderly and other marginalized groups, healthy community relations, community-based fact-checking, and education programs to counter hatred, violence and extremism."

This paper does not aspire to provide exhaustive solutions for addressing every aspect and actor involved in preventing and mitigating information operations in conflict scenarios. For example, the legal analysis highlights the pressing need to clarify and reinforce the application of international humanitarian law to safeguard civilians and ensure their access to crucial survival information. In this regard, the 2022 report, *Disinformation and*

freedom of opinion and expression during armed conflicts, by Special Rapporteur Irene Khan offers far more comprehensive recommendations. Another important entry point for international collective action is the United Nations Global Principles for Information Integrity presented by the UN Secretary-General António Guterres on June 24, 2024. In the words of Guterres, "These five principles — [societal] trust and resilience; independent, free, and pluralistic media; healthy incentives; transparency and research; and public empowerment — are based on an overriding vision of a more humane [information] ecosystem" (United Nations General Assembly 2022).

Nonetheless, in the present context, by building on its technical analysis and scenario planning, this paper aims to demonstrate that fostering new collaborations and adopting anticipatory, foresight-based methods will be essential to driving meaningful change, particularly as current threats in sectors such as AI, cyber security and biosecurity are still being governed in silos.

Inclusive Foresight for Better Prevention and Mitigation

Close, effective and sustainable partnerships between civilian private sector actors, technology leaders, civil society organizations, governments and military institutions should be convened to conduct combined foresight analyses across technological domains, including generative AI, cyber security and biosecurity. With an aim toward understanding the convergence of generative AI and other dual-use technologies with high-impact biological events, such groups could define a shared approach to prevention and mitigation. This paper's scenario demonstrates how disinformation campaigns about biological threats could be tailored to destabilize civilian populations and undermine trust in health systems and governing institutions, producing even more strategic effects and harmful impacts in situations of protracted conflict between two states.

It is urgent that governments collaborate with the private sector to create more efficient early warning systems to detect and analyze the sources, actors and modus operandi behind the information and psychological operations that target civilian populations. As shown in the paper's scenario, closer collaborations among policy makers, the defence sector, health-care providers, the commercial biotech industry and medical research institutions (including a network of expert scientists and health-care professionals for crisis mobilization) are crucial. These partnerships are essential for developing countermeasures to combat the information operations that might accompany or exploit the threat of a biological attack.

Foresight efforts should include cooperation with states affected by conflict: experts in conflict prevention should partner with private sector actors and civil society to better tailor prevention strategies to the specific threats and ethical needs of vulnerable communities. Such "inclusive foresight" could equip countries and agencies with the tools to articulate scenarios from which risk prioritization can emerge, particularly in conflict zones, as well as develop responsible approaches to leverage emerging technologies for prevention.

Works Cited

Agranovich, David. 2023. "Detect, Disrupt, Deter." In *Digital Front Lines: A sharpened focus on the risks of, and responses to, hybrid warfare,* a Special Report from FP Analytics with support from Microsoft, 32. https://digitalfrontlines.io/wp-content/uploads/sites/8/2023/08/digital-front-lines-report-FP-analytics-microsoft-2023.pdf.

Alfifi, Majid, Parisa Kaghazgaran and James Caverlee. 2018. "Measuring the Impact of ISIS Social Media Strategy." http://snap.stanford.edu/mis2/files/MIS2_paper_23.pdf.

Almohammad, Asaad. 2018. "ISIS Child Soldiers in Syria: The Structural and Predatory Recruitment, Enlistment, Pre-Training Indoctrination, Training, and Deployment." International Centre for Counter-Terrorism Research Paper. February 19. https://doi.org/10.19165/2018.1.02.

Baraniuk, Chris. 2018. "Surgeon David Nott: Hack led to Syria air strike." BBC, March 21. www.bbc.com/news/technology-43486131.

Beauchamp-Mustafaga, Nathan. 2024. "Exploring the Implications of Generative AI for Chinese Military Cyber-Enabled Influence Operations: Chinese Military Strategies, Capabilities and Intent." Testimony presented before the US-China Economic and Security Review Commission on February 1. RAND Corporation. www.rand.org/content/dam/rand/pubs/testimonies/CTA3100/CTA3191-1/RAND_CTA3191-1.pdf.

Bingle, Morgan. 2023. "What is Information Warfare?" The Henry M. Jackson School of International Studies, University of Washington. September 25. https://jsis.washington.edu/news/what-is-information-warfare/.

Bowcott, Owen. 2020. "UN warns of rise of 'cybertorture' to bypass physical ban." *The Guardian,* February 21. www.theguardian.com/law/2020/feb/21/un-rapporteur-warns-of-rise-of-cybertorture-to-bypass-physical-ban.

Brandt, Jessica. 2023. "Propaganda, foreign interference, and GenAI." Brookings, November 8. www.brookings.edu/articles/propaganda-foreign-interference-and-generative-ai/.

Burt, Tom. 2023. "The Face of Modern Hybrid Warfare." In *Digital Front Lines: A sharpened focus on the risks of, and responses to, hybrid warfare*, a Special Report from FP Analytics with support from Microsoft, 14–15. https://digitalfrontlines.io/wp-content/uploads/sites/8/2023/08/digital-front-lines-report-FP-analytics-microsoft-2023.pdf.

Byman, Daniel L., Chongyang Gao, Chris Meserole and V. S. Subrahmanian. 2023. *Deepfakes and International Conflict.* Foreign Policy at Brookings. January. www.brookings.edu/wp-content/uploads/2023/01/FP_20230105_deepfakes_international_conflict.pdf.

Carballo, Rebecca. 2023. "Using AI To Talk to the Dead." *The New York Times,* December 11. www.nytimes.com/2023/12/11/technology/ai-chatbots-dead-relatives.html.

Carter, Sarah R., Nicole E. Wheeler, Sabrina Chwalek, Christopher R. Isaac and Jaime Yassif. 2023. *The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe.* Nuclear Threat Initiative. October. www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/.

Concentric. 2019. "E-Recruits: How Gaming is Helping Terrorist Groups Radicalize and Recruit a Generation of Online Gamers." March 17. www.concentric.io/blog/e-recruits-how-gaming-is-helping-terrorist-groups-radicalize-and-recruit-a-generation-of-online-gamers.

Dorn, Sara. 2023. "Republicans Launch Eerie AI-Generated Attack Ad On Biden." *Forbes*, April 25. www.forbes.com/sites/saradorn/2023/04/25/republicans-launch-eerie-ai-generated-attack-ad-on-biden/.

Fecteau, Matthew. 2021. "The Deepfakes Are Coming." *War Room*, April 23. https://warroom.armywarcollege.edu/articles/deep-fakes/.

Federspiel, Frederik, Ruth Mitchell, Asha Asokan, Carlos Umana and David McCoy. 2023. "Threats by artificial intelligence to human health and human existence." *BMJ Global Health* 8 (5): e010435. https://doi.org/10.1136/bmjgh-2022-010435.

Fedorov, Mykhailo. 2023. "Lessons from Ukraine in the Heat of an Ongoing Hybrid War." In *Digital Front Lines: A sharpened focus on the risks of, and responses to, hybrid warfare,* a Special Report from FP Analytics with support from Microsoft, 12–13. https://digitalfrontlines.io/wp-content/uploads/sites/8/2023/08/digital-front-lines-report-FP-analytics-microsoft-2023.pdf.

Feldstein, Steven. 2023. "The Consequences of Generative AI for Democracy, Governance and War." *Survival* 65 (5): 117–42. https://doi.org/10.1080/00396338.2023.2261260.

Fixler, Annie. 2023. "Cyber-Resilience Helps Democracies Prevail Against Authoritarian Disinformation." In *Digital Front Lines: A sharpened focus on the risks of, and responses to, hybrid warfare,* a Special Report from FP Analytics with support from Microsoft, 11. https://digitalfrontlines.io/wp-content/uploads/sites/8/2023/08/digital-front-lines-report-FP-analytics-microsoft-2023.pdf.

FP Analytics. 2023. "Strategies for Reconciling International Humanitarian Law and Cyber Operations: A Q&A with Dr. Peter Maurer." In *Digital Frontlines: A sharpened focus on the risks of, and responses to, hybrid warfare,* a Special Report from FP Analytics with support from Microsoft, 28–29. https://digitalfrontlines.io/wp-content/uploads/sites/8/2023/08/digital-frontlines-report-FP-analytics-microsoft-2023.pdf.

Giles, Keir. 2023. "Russian cyber and information warfare in practice: Lessons observed from the war on Ukraine." Chatham House Research Paper. December 14. https://doi.org/10.55317/9781784135898.

Gisel, Laurent, Tilman Rodenhäuser and Knut Dörmann. 2020. "Twenty years on: International humanitarian law and the protection of civilians against the effects of cyber operations during armed conflicts." *International Review of the Red Cross* 102 (913): 287–334. https://doi.org/10.1017/S1816383120000387.

Gisselsson, David. 2022. "Next-Generation Biowarfare: Small in Scale, Sensational in Nature?" *Health Security* 20 (2): 182–86. https://doi.org/10.1089/hs.2021.0165.

Goel, Vindu and Shaikh Azizur Rahman. 2019. "When Rohingya Refugees Fled to India, Hate on Facebook Followed." *The New York Times,* June 14. www.nytimes.com/2019/06/14/technology/facebook-hate-speech-rohingya-india.html.

Guay, Joseph, Stephen Gray, Meghann Rhynard-Geil and Lisa Inks. 2019. *The Weaponization of Social Media: How social media can spark violence and what can be done about it.* Mercy Corps. November. www.mercycorps.org/sites/default/files/2020-01/Weaponization_Social_Media_FINAL_Nov2019.pdf.

Hart, Robert. 2022. "Clearview AI Fined $9.4 Million In U.K. For Illegal Facial Recognition Database." *Forbes,* May 23. www.forbes.com/sites/roberthart/2022/05/23/clearview-ai-fined-94-million-in-uk-for-illegal-facial-recognition-database/.

Hiebert, Kyle. 2024. "Generative AI Risks Further Atomizing Democratic Societies." Opinion, Centre for International Governance Innovation, February 26. www.cigionline.org/articles/generative-ai-risks-further-atomizing-democratic-societies/.

Hutchinson, William. 2006. "Information Warfare and Deception." *Informing Science: The International Journal of an Emerging Transdiscipline* 9: 213–23. https://doi.org/10.28945/480.

ICC. 2021. *Rome Statute of the International Criminal Court.* The Hague, The Netherlands: ICC. www.icc-cpi.int/sites/default/files/2024-05/Rome-Statute-eng.pdf.

ICRC. 2021. *Harmful Information – Misinformation, Disinformation and Hate Speech in Armed Conflict and Other Situations of Violence. ICRC Initial Findings and Perspectives on Adapting Protection Approaches.* July. https://shop.icrc.org/harmful-information-misinformation-disinformation-and-hate-speech-in-armed-conflict-and-other-situations-of-violence-icrc-initial-findings-and-perspectives-on-adapting-protection-approaches-pdf-en.html.

Jackson, Jasper, Lucy Kassa, Mark Townsend. 2022. "Facebook 'lets vigilantes in Ethiopia incite ethnic killing.'" *The Guardian,* February 20. www.theguardian.com/technology/2022/feb/20/facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing.

Jingnan, Huo. 2024. "How Israel tried to use AI to covertly sway Americans about Gaza," NPR, June 6. www.npr.org/2024/06/05/nx-s1-4994027/israel-us-online-influence-campaign-gaza.

Kastrenakes, Jacob. 2013. "Syrian Electronic Army alleges stealing 'millions' of phone numbers from chat app Tango." The Verge, July 22. www.theverge.com/2013/7/22/4545838/sea-giving-hacked-tango-database-government.

Katz, Eian. 2021. "Liar's war: Protecting civilians from disinformation during armed conflict." *International Review of the Red Cross* 102 (914): 659–82. https://doi.org/10.1017/S1816383121000473.

Khan KC, Karim A. A. 2023. "Technology Will Not Exceed Our Humanity." In *Digital Front Lines: A sharpened focus on the risks of, and responses to, hybrid warfare,* a Special Report from FP Analytics with support from Microsoft, 50–51. https://digitalfrontlines.io/wp-content/

uploads/sites/8/2023/08/digital-front-lines-report-FP-analytics-microsoft-2023.pdf.

Klepper, David. 2023. "Fake babies, real horror: Deepfakes from the Gaza war increase fears about AI's power to mislead." AP News, November 28. https://apnews.com/article/artificial-intelligence-hamas-israel-misinformation-ai-gaza-a1bb303b637ffbbb9cbc3aa1e000db47.

Lahmann, Henning. 2020. "Protecting the global information space in times of armed conflict." *International Review of the Red Cross* 102 (915): 1227–48. https://doi.org/10.1017/S1816383121000400.

MacDonald, Andrew and Ryan Ratcliffe. 2023. "Cognitive Warfare: Maneuvering in the Human Dimension." *Proceedings* 149 (4). The U.S. Naval Institute. www.usni.org/magazines/proceedings/2023/april/cognitive-warfare-maneuvering-human-dimension.

Marcellino, William, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao and Jackson Smith. 2023. *The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0: Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI.* RAND Corporation, September 7. www.rand.org/pubs/perspectives/PEA2679-1.html.

Marlatt, Greta E. 2008. "Information Warfare and Information Operations (IW/IO): A Bibliography." Monterey, CA: Dudley Knox Library, Naval Postgraduate School.

Marten, Kimberly. 2022. "Russia's Use of the Wagner Group: Definitions, Strategic Objectives, and Accountability." Testimony before the Committee on Oversight and Reform Subcommittee on National Security, United States House of Representatives, Hearing on "Putin's Proxies: Examining Russia's Use of Private Military Companies," September 15. https://docs.house.gov/meetings/GO/GO06/20220921/115113/HHRG-117-GO06-Wstate-MartenK-20220921.pdf.

Maslej, Nestor, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika et al. 2023. *Artificial Intelligence Index Report 2023.* Stanford, CA: Institute for Human-Centered AI, Stanford University. April. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

McGuire, Mike. 2021. "Nation States, Cyberconflict and the Web of Profit." *HP Threat Research Blog*, April 8. https://threatresearch.ext.hp.com/web-of-profit-nation-state-report/.

Meisenzahl, Mary. 2019. "ISIS is reportedly using popular Gen Z app TikTok as its newest recruitment tool." Business Insider, October 21. www.businessinsider.com/isis-using-tiktok-to-target-teens-report-2019-10?r=US&IR=T.

Morris, Tamer. 2024. "Israel-Hamas 2024 Symposium — Information Warfare and the Protection of Civilians in the Gaza Conflict." The Lieber Institute, West Point. January 23. https://lieber.westpoint.edu/information-warfare-protection-civilians-gaza-conflict/.

Mouton, Christopher A., Caleb Lucas and Ella Guest. 2023. *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach.* Research report, RAND Corporation. www.rand.org/pubs/research_reports/RRA2977-1.html.

Moy, Wesley R. and Kacper T. Gradon. 2023. "Artificial intelligence in hybrid and information warfare: A double-edged sword." In *Artificial Intelligence and International Conflict in Cyberspace,* edited by Fabio Cristiano, Dennis Broeders, François Delerue, Frédérick Douzet and Aude Géry, 47–74. Abingdon, UK: Routledge. https://doi.org/10.4324/9781003284093-4.

Mubarak, Rami, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dute, Saad Khan and Simon Parkinson. 2023. "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats." *IEEE Access* 11: 144497–529. https://doi.org/10.1109/ACCESS.2023.3344653.

Pandith, Farah and Jacob Ware. 2021. "Teen terrorism inspired by social media is on the rise. Here's what we need to do." NBC Think, March 22. www.nbcnews.com/think/opinion/teen-terrorism-inspired-socialmedia-rise-here-s-what-we-ncna1261307.

Pauwels, Eleonore. 2020a. *The Anatomy of Information Disorders in Africa: Geostrategic Positioning & Multipolar Competition Over Converging Technologies*. September 9. New York, NY: Konrad Adenauer Stiftung. www.kas.de/en/web/newyork/single-title/-/content/the-anatomy-of-information-disorders-in-africa.

———. 2020b. *Artificial Intelligence and Data Capture Technologies in Violence and Conflict Prevention: Opportunities and Challenges for the International Community.* Global Center on Cooperative Security Policy Brief. September. www.globalcenter.org/wp-content/uploads/GCCS_AIData_PB_H-1.pdf.

———. 2024. *Regulating the Role and Involvement of Offensive Proxy Actors in Cyberconflict.* March 27. New York, NY: Konrad Adenauer Stiftung. www.kas.de/en/web/newyork/veranstaltungsberichte/detail/-/content/regulating-the-role-and-involvement-of-offensive-proxy-actors-in-cyberconflict-1.

Permanent Mission of Liechtenstein to the United Nations. 2021. *The Council of Advisers' Report on the Application of the Rome Statute of the International Criminal Court to Cyberwarfare.* August. www.regierung.li/files/medienarchiv/The-Council-of-Advisers-Report-on-the-Application-of-the-Rome-Statute-of-the-International-Criminal-Court-to-Cyberwarfare.pdf.

Pomerantsev, Peter and Michael Weiss. 2014. *The Menace of Unreality: How the Kremlin Weaponizes Information, Culture and Money.* Special Report presented by The Interpreter, a Project of the Institute of Modern Russia. https://imrussia.org/media/pdf/Research/Michael_Weiss_and_Peter_Pomerantsev__The_Menace_of_Unreality.pdf.

Prier, Jarred. 2017. "Commanding the Trend: Social Media as Information Warfare." *Strategic Studies Quarterly* 11 (4): 50–85.

Raymond, Nate. 2022. "Facebook parent Meta to settle Cambridge Analytica scandal case for $725 million." Reuters, December 23. www.reuters.com/legal/facebook-parent-meta-pay-725-mln-settle-lawsuit-relating-cambridge-analytica-2022-12-23/.

Ridout, Brad, Melyn McKay, Krestina Amon and Andrew Campbell. 2019. *Mobile Myanmar: The impact of social media on young people in conflict-affected regions of Myanmar.* Yangon, Myanmar: Save the Children Myanmar and The University of Sydney. www.savethechildren.es/sites/default/files/imce/docs/mobile_myanmar_report_short_final.pdf.

Rodenhäuser, Tilman and Samit D'Cunha. 2023. "Foghorns of war: IHL and information operations during armed conflict." *Humanitarian Law and Policy* (blog), October 12. https://blogs.icrc.org/law-and-policy/2023/10/12/foghorns-of-war-ihl-and-information-operations-during-armed-conflict/.

Sandbrink, Jonas B. 2023. "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools." *arXiv*, June 24. https://doi.org/10.48550/arXiv.2306.13952.

Schafer, Bret, Nathan Kohlenberg, Amber Frankland and Etienne Soula. 2021. "Influence-enza: How Russia, China, and Iran Have Shaped and Manipulated Coronavirus Vaccine Narratives." Alliance for Securing Democracy. March 6. https://securingdemocracy.gmfus.org/russia-china-iran-covid-vaccine-disinformation/.

Segal, Dave. 2018. "How Bell Pottinger, P.R. Firm for Despots and Rogues, Met Its End in South Africa." *The New York Times*, February 4. www.nytimes.com/2018/02/04/business/bell-pottinger-guptas-zuma-south-africa.html.

Singer, Peter Warren and Emerson T. Brooking. 2018. *LikeWar: The Weaponization of Social Media.* Boston, MA: Houghton Mifflin Harcourt.

Spink, Lauren. 2023. *When Words Become Weapons: The Unprecedented Risks to Civilians from the Spread of Disinformation in Ukraine.* Center for Civilians in Conflict. October. https://civiliansinconflict.org/wp-content/uploads/2023/11/CIVIC_Disinformation_Report.pdf.

Stanham, Lucia. 2023. "Generative AI (GenAI) in Cybersecurity." Crowdstrike. November 26. www.crowdstrike.com/cybersecurity-101/secops/generative-ai/.

Stewart, Ruben and Georgia Hinds. 2023. "Algorithms of war: The use of artificial intelligence in decision making in armed conflict." *Humanitarian Law and Policy* (blog), October 24. ICRC. https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict/.

The UIC John Marshall Law School IHRC and Access Now, supported by Syrian Justice & Accountability Centre and MedGlobal. 2021. *Digital Dominion: How the Syrian Regime's Mass Digital Surveillance Violates Human Rights*. March. www.accessnow.org/cms/assets/uploads/2021/03/Digital-dominion-Syria-report.pdf.

United Nations Convention on Rights of a Child. 2021. General comment No. 25 on children's rights in relation to the digital environment. GC/25. March 2. www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation.

United Nations General Assembly. 2022. *Disinformation and freedom of opinion and expression during armed conflicts.* A/77/288. August 12. www.ohchr.org/en/documents/thematic-reports/a77288-disinformation-and-freedom-opinion-and-expression-during-armed.

United Nations Human Rights Commission. 2018a. *Report of the independent international fact-finding mission on Myanmar*. A/39/64. www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf.

———. 2018b. *Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar*. A/39/CRP.2. https://reliefweb.int/report/myanmar/report-detailed-findings-independent-international-fact-finding-mission-myanmar.

Ward, Antonia. 2018. "ISIS's Use of Social Media Still Poses a Threat to Stability in the Middle East and Africa." RAND, December 11. www.rand.org/blog/2018/12/isiss-use-of-social-media-still-poses-a-threat-to-stability.html.

Ward, Clarissa, Katie Polglase, Sebastian Shukla, Gianluca Mezzofiore and Tim Lister. 2020. "Russian election meddling is back — via Ghana and Nigeria — and in your feeds." CNN, April 11. https://edition.cnn.com/2020/03/12/world/russia-ghana-troll-farms-2020-ward/index.html.

Watts, Clint. 2022. "Preparing for a Russian cyber offensive against Ukraine this winter." *Microsoft On the Issues* (blog), December 3. https://blogs.microsoft.com/on-the-issues/2022/12/03/preparing-russian-cyber-offensive-ukraine/.

Whiskeyman, Andrew and Michael Berger. 2021. "Axis of Disinformation: Propaganda from Iran, Russia, and China on COVID-19." Policy Analysis. Washington, DC: The Washington Institute for Near East Policy. February 25. www.washingtoninstitute.org/policy-analysis/axis-disinformation-propaganda-iran-russia-and-china-covid-19.

Wirtschafter, Valerie. 2024. "The implications of the AI boom for non-state armed actors." Brookings. January 16. www.brookings.edu/articles/the-implications-of-the-ai-boom-for-nonstate-armed-actors/.

Yang, Zeyi. 2024. "Deepfakes of your dead loved ones are a booming Chinese business." *MIT Technology Review,* May 7. www.technologyreview.com/2024/05/07/1092116/deepfakes-dead-chinese-business-grief/.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* London, UK: Profile Books.

Zuidijk, Daniel. 2023. "Deepfakes in Slovakia Preview How AI Will Change the Face of Elections." Bloomberg. October 4. www.bloomberg.com/news/newsletters/2023-10-04/deepfakes-in-slovakia-preview-how-ai-will-change-the-face-of-elections.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org