PEACEKEEPING IN AN ERA OF CONVERGING TECHNOLOGICAL & SECURITY THREATS

Preventing Collective AI & Data Harms, Learning to Save Lives with Dual-Use Technologies

By Eleonore Pauwels

The field of peacekeeping is about to face an upheaval as it confronts renewed questions about its capacity to analyse multimodal and sensitive data about conflict situations, mitigate subsequent information risks, ensure accountability, and preserve public trust in an era of converging security threats. Threats to human rights and security triggered by the convergence of artificial intelligence (AI) with other dual-use technologies will require peacekeeping operations to anticipate new complex challenges in bridging the gap between early warning and response, and improving the safety and security of UN staff and of civilian populations.

Two complex matters of concern should feature high on the agenda of peacekeeping operations in the technological convergence era: first, how to protect civilian populations from collective AI and data-harms that could infringe on human rights, erode social cohesion and resilience, and impact future conflicts; second, how to develop internal capacity and muster cross-sector collaborations to mitigate the pervasive AI-led cyberthreats that could corrupt data-integrity in peacekeeping and seriously undermine trust in its mandate and operations. While challenges are rising when it comes to harnessing foresight, new skills and ensuring trust and accountability, peacekeeping actors also have a unique opportunity to approach technological and data-governance, including AI and cybersecurity, from a conflict-sensitive perspective.

Moving forward, the field of peacekeeping must exert robust normative leadership and strengthen a theory of no-harm, partnering with the next generation of civil society and private sector actors to protect civilian populations across conflict zones.

***

AI CONVERGENCE – POWER OVER POPULATIONS' BODIES AND MINDS

We have entered a technological era where our private and collective experience has become free material for behavioural surveillance.[1] Our "patterns of life" – our emotions and behaviours, our biometrics and biological (even neural)[2] data – can be turned into predictive insights to fuel epistemic and cyber-conflicts. This technical assemblage – termed "Internet of bodies and minds" – is born out of the convergence between AI, affective computing, and biotechnology. Combinations of dual-use technologies[3] are increasingly deployed in cyberspace, harnessed through critical infrastructure and industrial platforms. They empower some communities and disempower others, including in conflict-prone and fragile countries.

This convergence paradigm where AI acts as an innovation catalyst creates new knowledge and power asymmetries. For instance, data-collectors and data-brokers, including humanitarian organisations, hold unprecedented power over civilian populations, even more in a crisis region.

The increasing confluence of AI with other dual-use technologies is an epistemic revolution as much as a technological one. This epistemic revolution embodies new powerful methods and techniques to, not only analyse large swaths of data, but also manipulate the integrity of datasets and the functioning of algorithms. Algorithms can impact processes of knowledge-production in several distinct ways: 1) synthetizing new datasets from scratch by reproducing the characteristics of a certain type of information (from synthetizing individuals' faces, voice samples to proteins' structure); 2) corrupting the integrity and content of a digital trove of information (from altering medical scans, DNA sequences in genomes to satellite imagery used in situational awareness); and 3) manipulating the functioning, performance and predictive value of other algorithmic systems.

[1] Mirca Madianou, "The Biometric Assemblage: Surveillance, Experimentation, Profit, and the Measuring of Refugee Bodies," Television and New Media, 2 July 2019, https://journals.sagepub.com/doi/abs/10.1177/1527476419857682. See also, Pauwels E., 2019. "The New Geopolitics of Converging Risks". United Nations University. https://collections.unu.edu/eserv/UNU:7308/PauwelsAIGeopolitics.pdf

[2] Yuste, Rafael, et al. « Four Ethical Priorities for Neurotechnologies and AI ». Nature News, vol. 551, no 7679, November 2017. https://www.nature.com/news/four-ethical-priorities-for-neurotechnologies-and-ai-1.22960

[3] Dual-use technologies belong to a set of technologies that are conceived, designed, and deployed for beneficial purposes but can also inherently cause harm, either accidentally or as a result of deliberate, malicious intent.

These new AI-led techniques are becoming critical to how our societies produce and assess knowledge across disciplines and sectors. They are powerful game-changers for all fields involved in data-driven analysis, including peacekeeping. Under this paradigm, population datasets and their supporting digital and physical infrastructures are a global public good – and a growing target for data-manipulation and adversarial information operations. Such rising forms of data-manipulation compromise citizens' trust in the ability of government and the multilateral system more broadly to protect global populations from converging technological and security threats.

Recent studies in AI- and cyber-security demonstrated that algorithms can learn to manipulate the integrity of medical and genomics datasets, expanding a cyberattack's impact through health, biotech and biosecurity sectors. In 2018, researchers at Ben-Gurion University designed a malicious attack to modify cancer-data in hospital CT scans, generating false lung tumours that conformed to a patient's unique anatomy, leading to a misdiagnosis rate in excess of 90%.[4] In other words, they demonstrated how an attacker can corrupt the data-content of CT scans, using a deep-learning algorithm to inject or remove evidence of a medical condition. In 2019, researchers at Sandia National Laboratory used an autonomous malware to manipulate raw data within large curation of human genomes.[5] The malware was able to, not only compromise the functioning of a genetic analysis software, but also alter actual fragments of DNA sequences within individuals' genomes.

Both types of malicious data-tampering could result into misdiagnosis with impact on clinical decisions and potential lethal outcomes for patients. Even more sobering, by manipulating intelligence about infectious diseases in humans and pathogens, data-poisoning attacks could seriously undermine the integrity of the global knowledge-production cycle in biomedicine. **Relying on similar methods, adversarial attacks could destroy confidence in how peacekeeping generates analysis and intelligence about conflict situations.**

AN EPISTEMIC REVOLUTION FOR PEACEKEEPING

Across societies' analytical and data-driven efforts, the risk of adversarial information manipulation is to compromise and sow distrust in critical systems, from strategic knowledge, discourse and decision-making; research and industrial infrastructure; to security and governance mechanisms. This risk is amplified by the convergence of AI with other technologies and the subsequent interdependence of AI, cyber- and biosecurity domains: data-poisoning may soon infect country-wide genomics databases, and potentially weaponize biological research, nuclear facilities, manufacturing supply chains, financial trading strategies and political discourse. Unfortunately, most of these fields are governed in silos, without a good understanding of how dual-use technologies might, through convergence, create system-wide risks at a global level.

The rise of a new typology of converging security threats, such as data-manipulation attacks, is of high importance to peacekeeping for several reasons.

First, information manipulation already targets and will keep affecting populations in conflict-prone countries, generating civilian harm that goes beyond the expected humanitarian consequences of cyber-operations.[6] **Second**, such type of rising threats could contaminate all fields and sectors involved in data-driven analysis, including peacekeeping. The unprecedented capacity to determine and control information veracity and integrity can be weaponized to undermine situational awareness and security intelligence as well as trust and legitimacy in peacekeeping.

In the near- and long-term future, cyber- and epistemic conflicts – in which information veracity and data integrity is threatened – will merge with armed conflicts and civil wars. For instance, the uncontrolled spreading of manipulated media related to a ceasefire violation could increase the risk of conflict escalation. Modern conflicts increasingly involve use, misuse and abuse of populations' data, from breaches of confidentiality, behavioural surveillance, information disorders, to information infrastructure sabotage or

[4] Mirsky, Yisroel, et al. CT-GAN: Malicious Tampering of 3D Medical Imagery Using Deep Learning. January 2019. arxiv.org, https://arxiv.org/abs/1901.03597v3

[5] Corey M Hudson - From Buffer Overflowing Genomics Tools to Securing - DEF CON 27 Bio Hacking Village. www.youtube.com, https://www.youtube.com/watch?v=7du1TltZOJg

[6] Pauwels E., The Anatomy of Information Disorders, Konrad Adenauer Stiftung Foundation, 2020, ISBN: 978-3-95721-706-6, https://www.kas.de/en/web/newyork/single-title/-/content/the-anatomy-of-information-disorders-in-africa

disruption. These "collective data harms" are contributing to a new geopolitics of insecurity that cuts across societies and borders. Such threats are of particular concern in conflict-prone and conflict-affected countries due to weak regulatory frameworks and the growing cybersecurity and digital divide.

This paper aims to help peacekeeping actors, not only better anticipate and understand the current and evolving converging security threats landscape, but also reflect critically on the types of strategies, foresight methods, operational safeguards, human skills and technological capabilities needed in data-driven and future peacekeeping operations.

Section 1 maps the impact of technological convergence ("AI convergence") on the conflict landscape and related peacekeeping mandates. **Section 2** develops synopses and functional definitions of how AI and converging technologies can be misused in civilian conflicts. In particular, section 2 focuses on insecurity flashpoints when adversarial attacks target the integrity of populations' datasets and data-driven processes in critical information infrastructures and industrial systems. Section 2 also assess conflict parties' intent and capacities, and identify vulnerabilities in the modern supply chains of converging technologies used in conflict. **Section 3** provides two forward-leaning case-studies (threat-forecasting exercises) to illustrate how epistemic and cyber-conflicts will impact the future of peacekeeping. These case-studies aim to shed light on emerging and future vulnerabilities and insecurity flashpoints that could compromise the transition towards data-driven and, even predictive, peacekeeping. The first case-study is about the potential collective data-harms that could come from processing and managing online, large amounts of behavioural and contextual information about populations in a crisis region. The second case-study focuses on the potential for adversarial data-manipulation to derail the knowledge-production and intelligence life-cycle in peacekeeping. **Section 4** then reflects on the types of strategies, foresight methods, operational safeguards, human skills and technological capabilities that will be needed to confront, manage and mitigate converging AI and security threats. Section 4 also interrogates whether these skills/capabilities should be developed by UN peacekeeping actors or acquired through partnerships with the private sector or state bodies.

SECTION 1 – HOW THE CONVERGENCE OF AI AND OTHER DUAL-USE TECHNOLOGIES IMPACTS THE CONFLICT LANDSCAPE AND RELATED PEACEKEEPING MANDATE

Within national and international spheres, there is a lack of understanding of the threats that AI and converging technologies can pose at the individual human level, broader threats to populations, and geopolitical confrontations potentially triggered by combinations of technologies. These trends pose a particular governance and legitimacy challenge for the international community and an unforeseen burden on civilian populations affected by conflict.

The Evolving Conflict Landscape

Many landmark studies have documented with evidence the recent deterioration of the cyberthreat landscape[7] and the rising human cost of hostile cyberoperations.[8] Enhanced in their scope and sophistication, cyberattacks worldwide have not only targeted critical civilian sectors, from finance, health to energy, but also industrial control systems, nuclear power plants and complex supply chains, including in biotechnology.[9] The humanitarian sector is not immune to such rising threats as it relies on cyberspace for communication and logistical operations, including when bringing protection and assistance to victims of armed conflicts. The Covid-19 pandemic has also acted as a pressure to transfer humanitarian efforts "online," in cyberspace where security mechanisms and governance rules have to be strengthened.

[7] 2020 Report - Exploiting AI: How Cybercriminals Misuse and Abuse AI and ML – TrendMicro Research/Europol´S European Cybercrime Centre/UNICRI, https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/exploiting-ai-how-cybercriminals-misuse-abuse-ai-and-ml

[8] « The Potential Human Cost of Cyber Operations ». International Committee of the Red Cross, 20 June 2019, https://www.icrc.org/en/publication/potential-human-cost-cyber-operations

[9] USA, Cyberspace Solarium Commission Report, March 2020, https://drive.google.com/file/d/1ryMCIL_dZ30QyjFqFkkf10MxIXJGT4yv/view; Roadmap for Digital Cooperation: implementation of the recommendations of the High-level Panel on Digital Cooperation, Report of the Secretary-General, June 2020, https://www.un.org/en/content/digital-cooperation-roadmap/assets/pdf/Roadmap_for_Digital_Cooperation_EN.pdf

Exploiting the pandemics crisis, influence operations have contaminated social media networks to undermine public trust in important elements of civilian security, including scientific and policy emergency measures, as well as governance authorities critical to health, food, political, and economic stability.[10]

Rising Converging Security Threats: The increasing digital reliance of our societies creates a context of extreme fragility where converging technologies can be harnessed to amplify insecurity. What is substantially different in the current age of technological convergence is the potential for AI to weaponize the interdependence between crucial digital assets and security domains: growing multimodal and sensitive datasets collected about populations worldwide; data-analytics systems crucial to intelligence and governance (including in armed conflicts); and interconnected, automated industrial platforms and critical infrastructure.

Technological convergence therefore deeply impacts how resilient our societies will become to new forms of hybrid threats, connecting across security domains, merging civilian and military contexts, and at the boundary between war and humanitarian operations. In a nutshell, two defining socio-technical shifts will affect the future of conflicts:

▪ **Amplification of the Cyber-Threat Landscape by Autonomous Computing Systems**: The capacity of computing systems to develop autonomous behaviours will affect life and death scenarios in civilian contexts, outside of traditional military settings. Advances in AI can automate capacity for massive data-optimization, predictive intelligence, systems behavioural analysis, and anomaly detection. Relying on such functional augmentation, AI programs can enable autonomy in other technologies and industrial platforms (e.g., energy, food, medical and biotech sectors) that are critical to civilian populations' survival and well-being. The capacity of adversarial algorithms to manipulate automated protocols provides an increasing potential to weaponize smart civilian technologies, industrial control systems and manufacturing supply chains. In 2018, a petrochemical company with a plant in Saudi Arabia was targeted by a new kind of cyberattack, not designed to shut down operations, but to compromise its safety protocols and trigger an explosion.[11]

Due to the interconnectedness of cyberspace, adversarial manipulation of automated protocols and industrial safety protocols could lead to the subsequent shutdown of primary critical systems, from medical equipment, emergency communications, electric grid, levees and dams, to drinking water distribution and sewage management. Early warnings might not be detectable. The harm could be done remotely, on a large scale, and spill-over to essential humanitarian services provided to conflict-affected populations.

▪ **Targeting of Civilian Population Datasets and Information Infrastructure**: Converging technologies are merging the data of individuals' digital, physical, and biological lives, with potential for pervasive vulnerabilities. Targets in modern conflicts include not only civilian populations, but their personal data within critical information infrastructures. In 2019, during the conflict in Yemen, Houthi rebel groups have been vying for access to personal data of civilians that benefited from food distribution by the UN World Food Program.[12] Across societies and borders, hacks of electoral, medical, biometrics ID and social media datasets already have resulted in breaches of sensitive information, from ethnic backgrounds, health profiles to online behaviours.

In the absence of robust security frameworks, however, deep-learning algorithms can manipulate the content of population datasets, creating insecurity flashpoints and leading to widespread collective data harms, violations of the right to privacy, and compromised governance systems and data integrity crucial to health, food, and civilian security.[13] From confidential data on local informants to sensitive information on refugees and minority groups, the humanitarian cyber-environment processes complex and multimodal data, and presents a number of entry points for both, exfiltration and manipulation. Humanitarian actors

[10] "Covid-19, Crime Prevention and Criminal Justice Priorities: A Spotlight on Vulnerable Groups," UNICRI, December 2020, http://www.unicri.it/Publications/COVID-19-crimeprevention-vulnerable-populations

[11] Sanger, David E. « Hack of Saudi Petrochemical Plant Was Coordinated From Russian Institute ». The New York Times, 23 October 2018. NYTimes.com, https://www.nytimes.com/2018/10/23/us/politics/russian-hackers-saudi-chemical-plant.html

[12] Mark Latonero, "Stop Surveillance Humanitarianism," New York Times, 11 July 2019, https://www.nytimes.com/2019/07/11/opinion/data -humanitarian-aid.html

[13] Pauwels, 2019, UNU; Pauwels, 2020, KAS.

face unprecedented challenges to preserving the confidentiality, integrity and availability of multimodal population datasets and sensitive data-processes within critical systems and across assistance sectors.

Combinations of dual-use technologies pose unique threats to conflict-prone or conflict–affected regions and developing states. Such states are less able to prepare for low-probability, high-impact events such as a cyberattack on industrial control systems, and will generally be less resilient should one materialize. In a country where critical infrastructure is already under pressure, for instance, a cyberattack on hospital networks or the electrical grid could be devastating. Moreover, fragile states may be less prepared to deploy the critical expertise and enforce the global rules required to regulate converging technologies.[14] Ungoverned digital spaces could become havens for remote attackers with global reach. As a result, civilians and vulnerable populations could become common targets in modern tech-driven and urban warfare, which could turn more precise and more deadly.

Potential Impact on Peacekeeping Skills and Practices

The convergence of AI and other dual-use technologies will have strategic and long-lasting impact on the present and future of peacekeeping operations. These implications can be summarised under the below trends (and will be developed further in section 4):

▪ First, **when it comes to protection of civilian mandates (POC)**, UN peacekeepers may need to prepare and adapt for protecting populations from rising collective harms caused by the integration of AI in cyber- and information operations (cf. section 2 and 3). Such evolution of POC mandates, if support for it materialises, could raise crucial questions about the future normative and operational mechanisms that would help peacekeepers monitor (even anticipate) and report potential AI and cyberthreats to civilian populations.

▪ Second, **when it comes to its own capacity for data-driven and predictive analysis,** the field of peacekeeping needs to, not only work through its pervasive problems of information fragmentation and subsequent lack of trust in information processed at operational level,[15] but also prepare to mitigate potential threats to the integrity of its own datasets and analytical processes (cf. section 3). Effective mechanisms should secure the diverse digital repositories (from ad-hoc reporting systems in missions to more centralised efforts) that process sensitive data about situational awareness and populations' routine activities. Without robust (AI and cybersecurity) safeguards, data-driven peacekeeping could lead to unintended harm and erode public trust. Far beyond violations of privacy, unintended harm includes collective data breaches with serious security implications, especially when data is gathered from vulnerable populations. The UN has already been the target of offensive cyber-attacks,[16] and strong rules are needed to determine who will have access to sensitive information, how it will be stored, and what security measures will be used to ensure the integrity of the data.

▪ Third, while challenges are rising when it comes to harnessing foresight, new skills and ensuring data-security and accountability, peacekeeping actors have a unique opportunity to approach technological and data-governance, including AI- and cyber-security, from **a conflict-sensitive perspective**. There is an urgent need to strengthen a **theory of no-harm in the data and technological convergence space** and such effort would benefit from a conflict-sensitive, **operational understanding of how data permeates the socio-technical systems of conflict**.

Section 4 will cover specific recommendations on the types of technical skills, foresight methods, operational safeguards, and human and technological capabilities that might be needed to confront, manage and mitigate converging AI and cybersecurity threats that target information integrity.

[14] It is important to note that even conflict-prone and fragile countries face a proliferation of dual-use technologies. For instance, in several African countries, AI, data-capture and surveillance technologies are increasingly deployed through international economic and development programs such as China's Belt and Road Initiative. See, Pauwels, 2020, KAS.

[15] Sarah-Myriam Martin-Brûlé, "Finding the UN Way on Peacekeeping-Intelligence," International Peace Institute, April 2020, https://www.ipinst.org/wp-content/uploads/2020/04/2004-Finding-the-UN-Way.pdf

[16] Marion Laurence, What are the Benefits and Pitfalls of 'Data-Driven' Peacekeeping?, Policy Brief, Center for International Policy Studies, University of Ottawa, December 2019. https://www.cips-cepi.ca/wp-content/uploads/2020/01/policy-brief-marion-laurence-1.pdf

SECTION 2 – FUNCTIONAL MATRIX OF CONVERGING THREATS IN CONFLICT - HOW PARTIES TO CONFLICT ARE POTENTIALLY MISUSING CONVERGING TECHNOLOGIES?

Converging technologies are becoming complex, hybrid systems that are merging the data of our digital, physical and biological lives, with potential for pervasive vulnerabilities and emerging risks. In this context of convergence, the integration of AI within cyber- and information operations presents unprecedented challenges to preserving the confidentiality, integrity and availability of multimodal population datasets and sensitive data-processes within critical infrastructures and across sectors. The below **synopses (or "functional definitions")** forecast and explain how converging security threats materialise into potential civilian conflict scenarios.

▪ SYNOPSIS 1: COGNITIVE-EMOTIONAL CONFLICTS

Behavioural engineering leads to increased political polarisation, information disorders, social unrest, and violent ethnic conflicts.

Combined with facial and affect recognition, closed-circuit television cameras, and biometrics, AI is increasingly being used to profile people as they live, move, and feel.[17] Furthermore, AI systems can learn to interpret and predict individual human actions, as well as classify behaviours and emotions as "normal," "abnormal," or "harmful." Such detection capacity is also used for crowd analysis where AI can help predict crowd behaviours, map social interactions or grouping in crowds, and flag atypical behaviours. Converging technologies can therefore be harnessed to predict and engineer human behaviours, with potential for social and political control, with corrosive human rights implications. These implications include limits to self-determination and political agency, violations to privacy and data-protection, discrimination, and new forms of censorship in the virtual and physical civic space. For instance, behavioural surveillance could help states and non-state-actors alike anticipate populations movements during protests, elections, ceasefires, religious or social events, to better enforce repression.

What populations have to face in fragile context are new forms of data-predation and commodification that can spill-over, far beyond surveillance, to impact and disrupt democracies and elections. In countries where privacy and data protection policies are not translated into robust operational mechanisms, state and private sector actors can extract sensitive personal data from an array of online population databases for targeting ethnic and socio-economic groups.[18] They can exploit citizens' personal profiles and information networks for spreading rumours, targeted propaganda, hate speech, mis- and disinformation. Often, these narratives and falsehoods deliberately aim to stoke ethnic, religious, or political conflict. Political campaigns in Kenya in 2017 and in Nigeria in 2015 relied on video propaganda that built on ethnic and socio-economic tensions to target segments of the electorate defined by ethnicity, political leanings and age.[19] Such disinformation operations were also built on citizens' fears related to terrorism and public health crises. The rationale behind such sophisticated disinformation architecture is to immerse citizens in an alternative, virtual reality where they themselves become producers of emotional manipulation. Interestingly, this tactic muddies who is supposed to carry the burden of intent behind spreading hateful content.

The malicious manipulation of information is not a new phenomenon but the convergence of AI and behavioural data about populations is drastically upgrading methods, techniques and tools. With AI technologies that can generate forgeries and synthetize media from scratch (text, images, video and audio samples), the craft of emotional manipulation could become ever more powerful and cause harm to specific ethnic subgroups. The capacity of deep-learning algorithms to synthetise individuals' biometrics and behavioural data leads, not only to forgeries ("deepfake"), but also to a rising type of threats, "precision biometrics attacks." In 2018, IBM detected an AI malware that can hide a cyberthreat, such as WannaCry,

[17] Several countries in Sub-Saharan Africa face a proliferation of technologies that make bodies and minds increasingly traceable: for instance, the AI software called Sentry used to detect "abnormal behaviour" in the streets of Johannesburg; mobile biometric devices deployed by Uganda Police Force and that use AI to confirm a match on the spot; facial and behaviour recognition used in Zimbabwe to predict the movements and actions of people in public spaces. See, Pauwels, 2020, KAS.

[18] Muthuri R., Monyango F., and Karanja W., 2018. "Biometric technology, elections, and privacy: Investigating privacy implications of biometric voter registration in Kenya's 2017 Election Process." Centre for Intellectual Property and Information Technology Law. https://www.cipit.org/images/downloads/CIPIT-Elections-and-Biometrics-Report.pdf

[19] Pauwels, KAS, 2020.

in a video conference application and launch only when it identifies the face of the target.[20] Minorities could be stigmatized in new and powerful ways with cyberthreats "designed" to harm them. Another major concern is the potential for automated ethnic profiling or "techno-racism."[21] Drones and police body cameras equipped with facial recognition and other biometric-capture capabilities are increasingly used to profile participants in social and racial justice movements, even during peaceful demonstrations.[22] From privacy, agency to identity, fundamental rights of vulnerable groups will require adequate protection.

The reliance on AI to combine malicious synthetic media and predictive behavioural monitoring could also amplify information disorders. Such techniques of emotional manipulation could be used by factions in conflict, from ruling elites and political parties to terrorist groups.[23] In the near future, malicious actors will be able to rely on predictive behavioural analysis to identify the emotional triggers that push subgroups to violence. Social engineering, psychological manipulation, and other techniques of subversion and deception will be amplified. The mobilisation and polarisation of civil society could escalate into major social unrest and exacerbate clashes between factions involved in a conflict (for example, this could involve synthetic media reporting possible atrocity committed by representatives of disputing factions). As mentioned by Di Razza and Mamiya, "it is very possible, perhaps likely, that at least one such socio-political conflict rises to the level of mass atrocity against a civilian population within the next ten years."[24]

□ FOCUS ON ACTORS: CAPACITY ASSESSMENT

The behavioural surveillance of individuals and populations is a growing, decentralised industry and involves a set of techniques that can be outsourced and commoditized on demand. In a 2019 report on the impact of the targeted surveillance industry on human rights, the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression highlighted that, while States were largely responsible, companies appeared to be "operating without constraint" too, in a "free for all" private surveillance industry environment.[25] For years and in dozen African countries, corporations of lobbyists and data-brokers like the SCL Group have been analysing data about African populations, from health, nutrition, sanitation, weapons to militarized youth.[26] The behavioural surveillance industry has been functioning in the shadow of our digital economy for decades but the recent pandemic crisis provided a watershed moment for some governments across the globe to monitor citizens' routine activities under the guise of political legitimacy. Biometrics and personal data, movements and consumption patterns, conversations and behaviours, can increasingly be analysed by AI, sensing and data-capture technologies.

Converging technologies are also changing the strategic communications environments in which conflicts play out. Both state and nonstate actors can extremely easily feed their own narratives and mis- and disinformation to their constituents within and across borders. Relying on the aggressive campaigns generated by PR companies like Cambridge Analytica, domestic political parties in Africa have demonstrated increasing capacity and willingness to instrumentalise digital networks for inflaming existing racial, social and economic divisions between subpopulations.[27]

[20] « DeepLocker: How AI Can Power a Stealthy New Breed of Malware ». Security Intelligence, 8 August 2018, https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/

[21] In a June 2020 report, the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia, and related intolerance analysed different forms of racial discrimination in the design and use of emerging digital technologies. UN Human Rights Council, "Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis," A/HRC/44/57, 18 June 2020 (advance edited version).

[22] Malkia Devich-Cyril, "Defund Facial Recognition: I'm a Second-Generation Black Activist, and I'm Tired of Being Spied On by the Police, Atlantic, 5 July 2020, https://www.theatlantic.com/technology/archive/2020/07/defund-facial-recognition/613771

[23] Pauwels, KAS, 2020.

[24] Namie Di Razza and Ralph Mamiya, The Future of the Protection of Civilians in UN Peacekeeping Operations, https://peacekeeping.un.org/en/future-of-peacekeeping, p. 5.

[25] Human Rights Council, 2019. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". UN General Assembly. https://www.undocs.org/A/HRC/41/35

[26] Cadwalladr C., 2020. "Fresh Cambridge Analytica leak 'shows global manipulation is out of control". The Guardian. https://www.theguardian.com/uk-news/2020/jan/04/cambridge-analytica-data-leak-global-election-manipulation

[27] ''Digital Democracy, Analogue Politics: How the Internet Era Is Transforming Politics in Kenya''. Foreign Affairs. https://www.foreignaffairs.com/reviews/capsule-review/2019-08-12/digital-democracy-analoguepolitics-how-internet-era-transforming

When the Islamic State of Iraq and the Levant (ISIL) increased its power and visibility through social media, its extremely violent propaganda, including doctored videos, created a wave of online emotional warfare.[28] The violent anti-Islamic backlash that followed was then instrumentalized by ISIL for its recruiting strategies.

Since then, digital and emotional manipulation techniques, including the generation of synthetic media, have become both democratized and refined. In 2019, researchers in Israel published a new method for making so-called deepfakes by creating realistic face-swapped videos in real time, with no extensive facial data training. Deep learning algorithms can pinpoint facial biometrics features in a video, then align the source face to the target's face. Algorithms that do not need to be trained on each new face target provide a powerful toolkit to create realistic video forgeries at scale and with minimal know-how. In their article, the researchers warn about the potential for democratizing impersonations: "Our method eliminates laborious, subject-specific data collection and model training, making face swapping accessible to non-experts."[29] In the same vein, states and non-state violent actors could make use of other data-synthesis techniques, in particular for simulating voice-samples or generating false speeches and news articles. In 2019, engineers demonstrated that they needed only 13 hours and a few dollars to train algorithms to produce realistic UN speeches on a wide variety of sensitive topics from nuclear disarmament to refugees.[30]

AI research labs and private companies will need to think seriously about adequate safeguards when disseminating training data, source codes of concern, or information that could be harnessed to increase the ease of producing multimodal synthetic media. This became clear when, in 2019, the research institute OpenAI decided not to publish the tacit knowledge, training data and full algorithmic model of a new AI-powered text generator. Yet, in 2020, the same AI model (GPT-3) became publicly available, providing a platform powerful enough to create, from scratch, texts, arguments and opinionated book reviews that could be misinterpreted as coming from a human writer.[31]

The ability of converging technologies to monitor and then engineer human behaviour has direct implications for the UN's peacekeeping and human rights agenda. The next synopsis shows how the targeting of population datasets and digital knowledge infrastructure could increasingly be used by state and nonstate actors alike for adversarial purposes.

▪ SYNOPSIS 2: COLLECTIVE CIVILIAN AND INDUSTRIAL DATA-HARMS

AI-led cyberattacks could lead to data-manipulation that generate widespread civilian harms by both, corrupting societies' digital repositories, and compromising the functioning of critical infrastructure and industrial control systems.

Under technological convergence, the power of the state and private industry to impose a form of modern biopolitics – measuring individual and collective biology, neural[32] and genetic profiles – will keep expanding. Global responses to the COVID-19 pandemic have unveiled how sensors, algorithms, and computing power can be combined with biological data and used in technologies that monitor and optimize public health. And as the collection of biometrics, behavioural, genomics and neural data on populations accelerates, there is a greater need to be able to grasp and control the typology of datasets captured and commodified.

In the absence of solid information security measures, deep-learning algorithms can be misused to manipulate multimodal, sensitive population datasets, creating new insecurity flashpoints and leading to widespread collective data harms. As mentioned above, adversarial attacks on biomedical datasets have resulted in manipulation of sensitive information, from cancer data in patients' CT scans to the DNA

[28] Antonia Ward, "ISIS's Use of Social Media Still Poses a Threat to Stability in the Middle East and Africa," RAND, 11 December 2018, https://www .rand.org/blog/2018/12/isiss-use-of-social-media-still-poses-a-threat-to-stability.html. See Majid Alfif et al., "Measuring the Impact of ISIS Social Media Strategy," 2018, http://snap.stanford.edu/mis2/files/MIS2_paper_23.pdf

[29] Samantha Cole, "This Program Makes It Even Easier to Make Deepfakes," Vice News, 19 August 2019, https://www.vice.com/en_us/article/kz4amx /fsgan-program-makes-it-even-easier-to-make-deepfakes; Open Data Science, "FSGAN: Subject Agnostic Face Swapping and Reenactment," Medium, 30 September 2019, https://medium.com/@ODSC/fsgan-subject-agnostic-face-swapping-and-reenactment-2f033b0ea83c

[30] Bullock, J. & Luengo-Oroz M., Automated Speech Generation from UN General Assembly Statements: Mapping Risks in AI Generated Texts, https://www.unglobalpulse.org/wp-content/uploads/2019/06/1906.01946-1.pdf

[31] « OpenAI's New Language Generator GPT-3 Is Shockingly Good—and Completely Mindless ». MIT Technology Review, https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

[32] Yuste, 2017.

sequences of individuals' genomes. Scientists at Sandia National Labs have shown how algorithms could be used to automate data-poisoning attack with the intent to corrupt research intelligence within large curation of genomics data.[33] Such data-poisoning would not only affect how to detect and analyse pathogens. It could also corrupt the insights collected for decades on complex diseases, affecting researchers' capacity to target and treat specific afflictions in groups of patients. The adversarial techniques applied to the biotech and medical sectors can transfer to other data-driven domains.

These threats to population datasets can lead to research and economic sabotage and compromise data integrity crucial to health, food and civilian wellbeing. The most sobering implication with such new forms of data-poisoning is that they would radically undermine citizens' trust – trust in the accuracy of emergency data-systems, clinical-trials, medical counter-measures (such as vaccines and other therapeutic agents), and data-based research efforts. In turn, malicious actors – states and non-state actors alike – may seize this moment of radical "epistemic uncertainty" and "trust decay" for (cyber)criminal gains, competitive advantage in value and supply chains as well as for commercial and geostrategic influence.

Reaching across societies' analytical and data-driven efforts, adversarial information manipulation expands risks to the sabotage of critical infrastructure, industrial platforms, financial, security and governance systems.

Billions of people rely on the automated protocols that underpin most of our modern manufacturing platforms and industrial control systems. Experts have already warned against increased cyberattacks aimed at targeting safety systems that operate critical infrastructure such as electrical, water, and sanitation facilities.[34] Particularly in situations of armed conflict, the disruption of critical civilian infrastructure can have devastating effects for populations in the immediate and longer term and can also hamper humanitarian activities. Yet, the advent of automation provides an increasing potential to expand a cyberattack's impact through interconnected and cloud-based industrial sectors, including health and biotech industries.

In the near-term future, security experts are raising concerns about the risk of adversarial algorithms being used to automate cyberattacks on biotech manufacturing.[35] Malicious actors could corrupt networks of sensors to impact control decisions on biotech laboratories, and damage, destroy or contaminate vital stocks of vaccines, antibiotics, cell or immune therapies for cancer treatment. The combination of biological data-manipulation (related to both, humans and pathogens) and cyberattacks on biomanufacturing could have drastic economic consequences and lethal outcomes for populations. Such combination of events could escalate into biosafety and biosecurity risks.

Adversarial information operations that target the knowledge, industrial and governance sectors are a powerful type of hybrid threats. They may serve an array of offensive goals and involve broad coalitions of malicious actors, including states, non-state actors and surrogates. They target systemic vulnerabilities and different civilian and security interfaces, and interfere with multiple levels of strategic and emergency decision-making.

New forms of covert, adversarial data-manipulation are extremely hard to detect, creating new challenges for attribution. They go beyond causing severe civilian harms, and producing discrete safety and security incidents. What is potentially under attack is the data integrity and the robustness of our globalized intelligence and knowledge-production system. The result is not only to seriously erode a country's digital sovereignty, but also to undermine both, global leadership crisis response and populations' trust and resilience. The capacity of state and non-state actors alike to damage public confidence and destabilize critical governance institutions could have powerful, long-term implications for peace and security.

[33] Corey M Hudson - From Buffer Overflowing Genomics Tools to Securing - DEF CON 27 Bio Hacking Village. www.youtube.com, https://www.youtube.com/watch?v=7du1TltZOJg

[34] « The Potential Human Cost of Cyber Operations ». International Committee of the Red Cross, 20 June 2019, https://www.icrc.org/en/publication/potential-human-cost-cyber-operations

[35] Pauwels, UNU, 2019.

□ FOCUS ON ACTORS: CAPACITY ASSESSMENT

Most AI security studies have emphasized how data-poisoning attacks are relatively easy to engineer and do not require outstanding technical expertise.[36] Moreover, they exhibit a form of "interoperability" or "portability," as their inherent intrusive and deceptive mechanisms can transfer to many data-driven and computing domains. They could even harness, as a catalyst, the reliance of these domains on automation. The techniques used for injecting perturbational noise into hospital CT scans or genomics datasets require more sophistication and training than what would be needed to manipulate, or just even compromise, other types of datasets, from texts, numbers, instructions, pictures to datapoints in maps. Within cybersecurity frameworks, operational safeguards, however, would demand distinctive skills to implement data-authentication and verification mechanisms that could protect automated protocols.

To launch most types of adversarial data-manipulation, perpetrators would still need to infect the target datasets with malware. Yet, cybersecurity policies of healthcare, industrial and even humanitarian systems, have mostly focused on issues of data confidentiality and availability, but not necessarily data integrity.[37] Progressive automation of analytics and industrial systems may augment level of interconnectedness and vulnerability to AI-led cyberattacks. Data-repositories, which are not directly connected to the Internet, may still be indirectly connected via the facility's internal network. They are also vulnerable to social engineering attacks, physical access and insiders. For instance, cyber criminals and nation-state actors have already mounted targeted cyber-operations against biotech firms researching, producing and distributing Covid-19 vaccines. In December 2020, IBM researchers and the US Cybersecurity and Infrastructure Security Agency (CISA) unveiled global social engineering attacks "intended to steal the network log-in credentials of corporate executives and officials at global organizations involved in the refrigeration process necessary to protect vaccine doses."[38] The underlying goal could be to access and manipulate shared information about how the vaccine is shipped, stored, kept cold and delivered.

For a host of reasons, adversarial data-manipulation could be performed by malicious actors, with relatively sophisticated AI and cybersecurity skills, and could increasingly integrate the offensive toolkit of states, cybercriminals and potentially non-state violent actors. As well documented in the ICRC's report on the potential human cost of cyberoperations:

"Cyber tools and methods can proliferate in a unique manner that is difficult to control. First, cyber space is a global domain: provided that the attacker can overcome the cyber security and defence measures in place, any network node and information residing on the network can be accessed from anywhere in the world. At the same time, cyber tools can be repurposed or reengineered. The combination of these two characteristics means that when cyber tools have been used, stolen, leaked or otherwise become available, actors other than those who developed them might be able to find them, reverse engineer them, and reuse them for their own purposes."[39]

When it comes to hostile cyberoperations, proliferation of tools and tacit knowledge should raise concerns and it is likely that cyber criminal groups and private firms will eventually offer as a paid service to wage adversarial attacks on the analytics architecture and industrial control systems that support most infrastructures critical to civilian populations.

VULNERABILITIES IN TECHNOLOGIES' DECENTRALISED SUPPLY CHAINS

As they converge, technologies become more automated and decentralized, blurring who is responsible for technologies' misuses, leading to what the author calls "atomized responsibility and liability." Technological convergence and its implications are rarely understood when private sector actors define technical and normative standards. Furthermore, ethical priorities of populations affected by violent conflict are often

[36] Mirsky, Yisroel, et al. CT-GAN: Malicious Tampering of 3D Medical Imagery Using Deep Learning. janvier 2019. arxiv.org, https://arxiv.org/abs/1901.03597v3; Finlayson, Samuel G., et al. « Adversarial Attacks on Medical Machine Learning ». Science (New York, N.Y.), vol. 363, no 6433, mars 2019, p. 1287-89; Allyn, Jérôme, et al. « Adversarial attack on deep learning-based dermatoscopic image recognition systems ». Medicine, vol. 99, no 50, December 2020.

[37] See Mirsky, 2019.

[38] Sanger, David E., & Sharon LaFraniere. « Cyberattacks Discovered on Vaccine Distribution Operations ». The New York Times, 3 December 2020. NYTimes.com, https://www.nytimes.com/2020/12/03/us/politics/vaccine-cyberattacks.html

[39] ICRC, 2019, p 7.

absent from self-regulation and corporate normative principles. These principles need to be translated and turned into viable normative practices that are "conflict-sensitive"[40] and can be overseen and tested for transparency and accountability.

Decisionmakers in the public and private sectors will therefore have to adapt, revise, and upgrade governance models to mitigate harms to populations affected by conflicts in an era of technological decentralization and automation, in particular when responsibility and liability for harms are not clearly defined and accounted for. As well explained by experts in conflict prevention,[41] the next decade is likely to witness an increase in "grey zone" conflicts – competitions among and within powerful tech-leading states and non-state actors that operate under the threshold of war, but may cause severe civilian harms. Regional powers like China and the U.S. have spent decades acquiring requisite technological and human capital in both genomics and AI, and have begun competing over digital assets. Rising geopolitical tensions already revolve around the commodification of a new resource emerging from the convergence of the AI and biotech industries: populations' biological and genomic data. Tech-leading states increasingly seek strategic positioning to upgrade their technological convergence capital and build and control the digital roads of cyberspace.

In this context, there is an urgent need to anticipate and devise how the private and public sectors will harness and regulate technologies' dual-use potential. More sobering, it will become pressing to monitor how authoritarian and non-authoritarian regimes alike will be able to co-opt powerful private sector capabilities generated by technological convergence for population surveillance, adversarial attacks, data-predation, power- and resource- capture. The complex and decentralised supply chains of converging technologies could lead to an unprecedented diffusion of power and dual-use potential in conflicts.

SECTION 3 – CASE-STUDIES: PEACEKEEPING IN AN ERA OF AI AND CYBERTHREATS

The field of peacekeeping faces an unprecedented challenge when it comes to its capacity to both, better harness powerful converging and predictive technologies to preserve peace, and protect vulnerable populations from collective data-harms in crisis regions. Essentially, the field is learning to modernise and save lives with dual-use technologies.

Peacekeeping is already a field that uses data-capture technologies and intelligence collection to map and understand recurrent conflict patterns and forecast potential crises.[42] UN peacekeepers need to integrate converging technologies to digitize, share, and secure the information they collect from open sources, human informants, and data-capture techniques. They also need to monitor and record online how armed nonstate actors evolve and blend into civilian environments, collude with transnational criminal networks, and adapt their attack strategies to new domains, including cyberspace. UN experts must scrutinize how hate speech and incitements to violence contaminate the lifeblood of social media and private messaging applications in countries where ethnic and socioeconomic tensions prevail.

The peacekeeping cyber environment therefore processes extremely sensitive datasets. And, while the field is progressively upgrading data-literacy skills and analytical processes, certain parts of its intelligence ecosystem suffer serious fragmentation and an endemic lack of trust by units on the ground operations. This latent degradation of confidence is an important vulnerability to the epistemic threats posed by AI convergence and, at the same time, a call and a catalyst for improvement and modernisation.

Below, two forward-leaning case-studies (or threat-forecasting exercises) aim to shed light on emerging vulnerabilities and insecurity flashpoints that could compromise the transition towards data-driven and, even predictive, peacekeeping. The first case-study is about the potential collective data-harms that could come from processing and managing online, large amounts of behavioural and contextual information

[40] JustPeace Labs, Technology in Conflict: Conflict Sensitivity for the Tech Industry, 2020. https://justpeacelabs.org/wp-content/uploads/2020/07/JustPeace-Labs-Conflict-Sensitivity-for-Tech-Industry.pdf

[41] See excellent papers by Adam Day and Paul Williams on the Future of Peacekeeping Operations, https://peacekeeping.un.org/en/future-of-peacekeeping

[42] See Olga Abilova and Alexandra Novosseloff, "Demystifying Intelligence in UN Peace Operations: Toward an Organizational Doctrine," International Peace Institute, June 2016, https://www.ipinst.org/wp-content/uploads/2016/07/1608_Demystifying-Intelligence.pdf. See also Martin-Brûlé, 2020 and Laurence, 2019.

about populations in crisis region. The second case-study focuses on the potential for adversarial data-manipulation attacks to derail the knowledge-production or intelligence life-cycle in peacekeeping.

CASE-STUDY 1: Peacekeeping and the "Internet of Bodies and Minds"

The field of peacekeeping increasingly monitors, records and analyses "patterns of life" about civilian populations living in fragile countries or those prone to violent outbreaks.[43] The analysis of populations' routine activities serves to predict violence drivers and patterns (e.g., sustained human rights violations and online hate speech targeted at ethnic subgroups) and early-warning signals of impending crises (e.g., changes in social media or city traffic, movements of refugees or armed groups) or to identify violent nonstate actors by identifying distinct features of the activities of a specific group. Other sources include communications metadata and internet connection records, but also extend to location and activity tracking, financial transactions, and social media activity.

Beyond social media and open-source intelligence, peacekeeping actors collect growing amounts of behavioural information on populations using automated data-capture technologies. Access to the large data sets captured through satellites and drones equipped with video surveillance is translated into dashboards and digital maps generated for improving situational awareness, locating outbreaks of violence, monitoring the position and movements of troops and identifying civilian infrastructure in need of protection. Such "risk maps" can visualise the probability of encountering checkpoints controlled by child soldiers in specific areas. They can reveal the location, settlements, and movements of ethnic subgroups, minorities, and refugees. These risks have implications for peacekeeping operations but are also heightened by the increasing permeability of critical datasets collected in humanitarian operations.[44] One example includes the biometric data from Syrian refugees that are systematically collected to create a form of "cross-border identity" in complicated displacement situations.[45] "[O]fficials providing medical aid to Syrian refugees in Greece were so concerned that the Syrian military might exfiltrate information from their database that they simply treated patients without collecting any personal data."[46]

UN agencies and humanitarian actors are increasingly reliant on the capabilities of digital platforms and private sector leaders in the field of AI, predictive data analytics, and biometric identity management systems.[47] And there are signs that a nascent assemblage of AI and data-capture technologies might be migrating to the conflict analysis spheres.[48] In the near-future, the vast amount of digital information and routine behaviours generated by populations could be analysed through AI-led computing. Algorithms are gaining access to what the author calls "the Internet of bodies and minds," a high volume and rich variety of behavioural data sets collected by smart sensing technologies in mobile devices and within cities' infrastructures.[49]

Troves of digital information about individuals' locations, livelihoods, behaviour, and opinions can potentially be subject to an array of powerful surveillance practices by malicious actors.[50] Population subgroups could be targeted by states or violent non-state actors for the information they share online

[43] UN Peacemaker Digital Toolkit, https://peacemaker.un.org/digitaltoolkit ; UN Department of Political and Peacebuilding Affairs (UNDPPA), "E-Analytics Guide: Using Data and New Technology for Peacemaking, Preventive Diplomacy and Peacebuilding," 2019, https://beta.unglobalpulse.org/wp-content/uploads/2019/04/e-analyticsguide2019.pdf ; Allard Duursma and John Karlsrud, "Predictive Peacekeeping: Strengthening Predictive Analysis in UN Peace Operations," Stability Journal, 13 February 2019, https://www.stabilityjournal.org/articles/10.5334/sta.663 ; Weisi Guo, Kristian Gleditsch, and Alan Wilson, "Retool AI to Forecast and Limit Wars," Nature, no. 562 (October 2018), pp. 331–333, https://www.nature.com/articles/d41586-018-07026-4

[44] In 2019 the World Food Programme (WFP) launched a five-year engagement with Palantir to optimize food delivery to populations on a biometric registry. The same year, WFP started a global algorithmic monitoring project to map signs of food insecurity using technology expertise developed by Microsoft, Google, and Amazon.

[45] Madianou, "Biometric Assemblage."

[46] Latonero, NYT, 2019.

[47] WFP, "Palantir and WFP Partner to Help Transform Global Humanitarian Delivery," 5 February 2019, https://www.wfp.org/news/palantir-and-wfp -partner-help-transform-global-humanitarian-delivery

[48] See Note 43

[49] Jay Stanley, "The Dawn of Robot Surveillance: AI, Video Analytics, and Privacy," American Civil Liberties Union, 2019, https://www.aclu.org/sites /default/files/field_document/061819-robot_surveillance.pdf

[50] International Committee of the Red Cross (ICRC), "Artificial Intelligence and Machine Learning in Armed Conflict," 6 June 2019, https://www.icrc .org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach; Sharon Weinberger, "Private Surveillance Is a Lethal Weapon Anybody Can Buy," New York Times, 19 July 2019, https://www.nytimes.com/2019/07/19/opinion/private-surveillance-industry.html

about hate speech, violence, election fraud, the activities of armed groups, and impending attacks. UN staff 's reliance on social platforms and digital channels also creates abundant metadata that can be used to profile and physically attack ethnic subgroups.[51] Such digital profiles can be exploited for surveillance, harassment, and can endanger the safety of local informants.[52] Local informants are prime targets for retaliation if they are known to provide security forces with crucial information about the location and movements of armed groups, as well as insights into the funding and criminal activities of transnational violent extremist groups.[53]

In the near-future, monitoring and reporting activities may increasingly integrate AI and data capture technologies, and face heightened cybersecurity risks. In the aftermath of the COVID-19 crisis, an array of processes for population monitoring and tracking have become digital, augmented by AI and sensing technologies. In the same vein, data optimization and predictive tools for situational awareness, social media, and behavioural analysis have the potential to modernize the field of peacekeeping. Moving forward, peacekeeping actors will need to thoroughly secure the behavioural and contextual information they collect about populations. Protecting such sensitive data sets from digital manipulation and cyber-exfiltration will be a complex challenge, but it is crucial to ensure the security of civilians and critical information infrastructure. A cyberattack by a state or violent nonstate actor could potentially exfiltrate sensitive information for surgical offensives or strikes. Recent deconfliction attempts revealed the threat of data misuse as discussed during a briefing to the Security Council on the humanitarian situation in Syria. In 2019, six different attacks targeted deconflicted civilian locations and humanitarian movements in northwest Syria. The Under-Secretary-General for Humanitarian Affairs and Emergency Relief Coordinator shared with the Security Council his conclusion "that in the current environment deconfliction is not proving effective in helping to protect those who utilize the system."[54]

In fragile or conflict settings, civilians are at risk of cyber- and physical attacks if they are targeted by social and emotional engineering tactics. As far back as 2013, social engineering attacks by pro-government electronic actors used the strategic interests and weaknesses of Syrian opposition activists.[55] Exploitation of breaches in humanitarian and peacekeeping datasets could lead to targeted forms of emotional engineering and behavioural control. AI malware can learn to watch, track, and evaluate individuals' emotions, language, and behaviour, impersonating trusted contacts within professional and personal networks, making communications generated by AI malware almost indistinguishable from human peer communications. An AI system that has been taught to study the behaviour of social network users and implement spear-phishing attacks on them has been able to perform more than six times as efficiently as humans and with a higher conversion rate.[56] AI malware could be trained to impersonate a strategic contact in a trusted network or to perform "precision biometrics attacks" by tailoring their offensive strategies to the facial features of the human targets they need to manipulate. Potential targets could be individuals within networks of civil informants, or UN staff and peacekeepers. In the humanitarian cyber-environment, the vulnerability of digital identity profiles and biometric databases will remain a constant and long-term concern.

Under the convergence of AI and other converging technologies, data harms are "collective" and invasive as population datasets are increasingly connected and a growing target for exfiltration and manipulation. Such vulnerabilities could therefore have corrosive implications, undermining confidence in the neutrality,

[51] Privacy International, "Doing No Harm in the Digital Era," 11 December 2018, p. 17, https://privacyinternational.org/report/2509/humanitarian -metadata-problem-doing-no-harm-digital-era

[52] Delphine van Solinge, "Digital Risks for Populations in Armed Conflict: Five Key Gaps the Humanitarian Sector Should Address," Humanitarian Law and Policy, 12 June 2019, https://blogs.icrc.org/law-and-policy/2019/06/12/digital-risks-populations-armed-conflict-five-key-gaps-humanitarian -sector

[53] Human Rights Watch, "By Day We Fear the Army, by Night the Jihadists," Human Rights Watch, 21 May 2018, https://www.hrw.org/report/2018/05 /21/day-we-fear-army-night-jihadists/abuses-armed-islamists-and-security-forces. See also Karlsrud, 2019 and Laurence, 2019.

[54] Mark Lowcock, "Briefing to the Security Council on the Humanitarian Situation in Syria," UN Office for the Coordination of Humanitarian Affairs, 30 July 2019, ERC_USG Mark Lowcock Statement to the SecCo on Syria- 30July2019 - as delivered.pdf (reliefweb.int)

[55] John Scott-Railton and Morgan Marquis-Boire, "A Call to Harm: New Malware Attacks Target the Syrian Opposition," Munk School of Global Affairs Research Brief, no. 19 (June 2013), https://paper.seebug.org/papers/APT/APT_CyberCriminal_Campagin/2013/19-2013-acalltoharm.pdf.

[56] John Seymour and Philip Tully, "Weaponizing Data Science for Social Engineering," Black Hat, https://www.blackhat.com/docs/us-16/materials/us -16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf

impartiality, independence and legitimacy of peacekeeping mandates. Yet, the most enduring harm would be on civilian trust – trust in peacekeeping and humanitarian assistance in general. For UN agencies, it is of the utmost urgency to determine the right balance between proportionality in data collection, diligent sharing policy, and effective mechanisms for securing population data sets. This imperative will only grow stronger if peacekeeping actors progressively adopt the deployment of AI and sensing technologies for situational awareness, population monitoring, and digital investigations.

Matters of concern expand beyond traditional requirements in cyber and data-security. The field of peacekeeping needs to work through its pervasive problems of information fragmentation and subsequent lack of trust in the intelligence processed at operational level. As explained in the next case-study (2), peacekeeping operations should also be better prepared to mitigate potential threats to the integrity of its own datasets and analytical processes.

CASE-STUDY 2: Adversarial Data-Manipulation and Predictive Peacekeeping

The field of peacekeeping faces significant challenges with regard to its capacity for data analysis, anticipation, and decision-making in times of crisis. The increasing use of technology to collect situational awareness and population data comes with a greater need for analytical capabilities to transform this intelligence into planned strategies and bridge the warning-response gap. Some parts of the UN peacekeeping system already appear to be suffering from “sensory overload.” They lack the capacity to sift through the massive amounts of information generated by social media analysis, intelligence collection and data-capture technologies.[57] The field of peacekeeping could therefore be tempted to accelerate the adoption of automated algorithmic systems for managing the data flood and improve its capacity for predictive and diagnostic analysis.

Several centralised efforts are focused on the predictive analysis and forecast of strategic elements of situational awareness.[58] Launched in 2014, the Situational Awareness Geospatial Enterprise (SAGE) is a web-based database that allows peacekeepers to record relevant incidents (including violent outbreaks, human rights abuses, troop movements, abductions, protests, livestock theft) under the form of structured data (including event type, location, number and ethnicity of victims, and affiliation of perpetrators). Joint Mission Analysis Centres (JMACS) are leading to a more integrated and predictive approach to data-driven peacekeeping with detailed reporting, scenario-based papers and warning notes. The Comprehensive Performance Assessment System (CPAS) manages data-analytics about performance and accountability of peacekeeping missions. An array of digital maps (for example, Ushahidi and ACLED) and analytical systems (such as Crimson Hexagon/Brandwatch) also play a major role in helping peacekeepers monitor real-time developments on the ground, including violent incidents, areas of control, the position and movements of troops, and population movements.

Just like data protocols in industrial control systems, genomics software analysis, collection of biomedical data, and hospital CT scans, most of these datasets and predictive analytics in peacekeeping are potential targets for adversarial data-manipulation.

As the combination of AI and big data-capture (including satellites and GIS) becomes essential to situational awareness, adversarial attacks could compromise data-driven analysis in peacekeeping. Manipulation of existing datasets like SAGE, even combined with the injection of synthetic data (ranging from text/news reports, visual or audio evidence), could compromise signals’ reporting and situational awareness analysis. Within SAGE for instance, a malware could be trained to manipulate the content of reported incidents (modifying words that label types of event, locations, perpetrators’ affiliation, victims’ ethnicity, reported numbers of armed groups involved, and threat-identification codes). A malware could also learn to inject false elements (such as erroneous spatial coordinates, number of soldiers, or forged visual/audio/video samples) into recorded events. The resulting harm would be to steer peacekeeping missions into wrongly prioritizing certain affected populations over others, operating in particular areas over others, or otherwise taking decisions detrimental to the neutrality impartiality, and independence of their mandate.

[57] Abilova and Novosseloff, “Demystifying Intelligence in UN Peace Operations.”

[58] See Karlsrud 2019 and Laurence 2019.

Without robust integrity safeguards, applying machine learning to SAGE with the aim to turn current analytical processes into predictive peacekeeping could amplify the potential for adversarial attacks. Malicious actors may use generative adversarial networks to, not only tamper with the accuracy and integrity of datasets used to train predictive algorithms in SAGE, but subsequently interfere with the outcome of the analysis and decision-making. In other words, adversarial data-manipulation could be designed to corrupt the predictive ability of machine-learning algorithms to identify threats based on monitoring and reporting activities. Corrupting the integrity of incident reports could drastically undermine the confidence level of predictive algorithms when they aim to ascribe threat potential to a combination of variables and incidental patterns.

The deployment of AI-enabled data-manipulation will drastically alter the relationship to evidence and truth across many peacekeeping activities, including conflict investigations, ceasefire supervision and supporting verification mechanisms, observation, monitoring and reporting. The capacity of a range of actors to influence conflict analysis, peacekeepers' perceptions, and public opinion with misleading information could have powerful long-term implications for the role of the United Nations in maintaining peace and security. Such unprecedented capacity to determine and control information veracity and integrity can be weaponized to undermine contextual and security intelligence as well as trust and legitimacy in peacekeeping.

The potential threats from adversarial data-manipulation are amplified by a pervasive lack of a rigorous and standardised approach to managing fragmented data-collection and curation efforts in peacekeeping. A 2020 IPI report crucially points to duplication of official, unofficial and ad-hoc information reporting, lack of robust data security practices and lack of trust in the intelligence collected by different operational components on the ground. As underlined by Martin-Brûlé, "in addition to these official databases [I2, SAGE or Cosmos], almost every unit, including the U2, police, JMAC, JOC, and human rights division, has its own unofficial databases; this multiplication of databases results from a lack of trust in the platforms."[59]

One major issue with securing situational awareness and peacekeeping data is therefore that its underlying digital infrastructure is relatively fragmented. It operates with different data-protection safeguards and is built on a mix of expert-based centralised efforts, ad-hoc reporting and, even, open-access environment. This means that many of the standard protocols, software, and best practices need to account for potential risks and misuse, such as data-exfiltration and manipulation. Protecting peacekeeping data requires attention to several key issues:

(a) Securing the data-infrastructure but also the data-capture, storage and validation equipment: Including mobile phones, web-cameras, full-motion video capture, ground-based sensors, unarmed unmanned aerial vehicles, geographic information systems, etc.

(b) Assuming that software pipelines (which may contain a couple or dozens of individual programs) have not had a formal security audit, and that commercial solution packages may use open-source software.

(c) Understanding that data collected or processed through a third-party service provider may have been handled on cloud servers (including foreign servers), which could constitute a threat to data integrity.

(d) Maintaining the integrity of data-in-motion and data-at-rest and, therefore, preserving integrity over the life-cycle of datasets—from collection, curation, processing, analysis, and long-term storage. A host of techniques exist to help secure data integrity, and, if this point seems trivial, it is important to understand that fragmented cyber- and data-ecosystems do not necessarily enable the most updated techniques. For instance, encryption can be used to protect data-at-rest. Secure multiparty computation[60] can help protect data-in-motion. Data authentication and verification

[59] See Martin-Brûlé, p. 16.

[60] Protocols for secure multiparty computation (MPC) enable a set of parties to interact and compute a joint function of their private data inputs while revealing nothing but the output.

mechanisms, such as cryptographic checksums,[61] are critical to ensure data integrity. Another method for detecting adversarial data-manipulation is digital watermarking.[62] Tampered images can also be spotted with anomaly-detection algorithms.

(e) Appreciating the dual-nature of situational awareness and intelligence collection, and assuming that all data providing situational awareness or insights about populations' routine activities (no matter how fragmented) contains information that can be used to target vulnerable individuals, subgroups and on-the-ground operations.

Protocols, processes, and security plans must be established in order to provide access to peacekeeping data by authorized parties and limiting access only to those individuals. Modelling and simulating the ways in which peacekeeping data are stored, accessed, and retrieved for analysis is a useful method for testing such data systems, forecasting potential threats, identifying systemic vulnerabilities, and building solutions and mitigation plans to address them. Increasingly, diverse sectors facing information security risks – for instance, in genomics and biomedicine – rely on these forms of sandboxing or operational foresight. In section 4, this paper will focus on crucial skills/capabilities, but also opportunities for the field of peacekeeping to practice combined foresight analysis across security domains and sectors, learning from other fields (such as cyber- and genomics security).

***

In a nutshell, the two case-studies have demonstrated how monitoring and reporting activities, including situational awareness analysis and intelligence collection, are rapidly becoming an element of UN peacekeeper mandates in which technological and data-governance will have important ramifications.

SECTION 4 – PREPARING FOR THE FUTURE – FORESIGHT STRATEGIES, OPERATIONAL SAFEGUARDS, HUMAN SKILLS AND TECHNOLOGICAL CAPABILITIES

The field of peacekeeping will need to develop the skills, capacity and the operational and normative frameworks required to maintain peace in the cyber domain. Substantial upgrade in professionalisation and human skills, technological capacity and data-governance will be strategic to the evolving peacekeeping mandate. In particular, this paper has emphasized three areas where preparedness, advanced technological capabilities and normative leadership will be critical:

▪ First, in regard to **POC mandate**, UN peacekeepers will need to prepare and adapt for protecting populations from rising collective harms caused by the integration of AI in new types of adversarial cyber- and information operations. In this context, the role of UN peacekeepers could range from **monitoring, reporting and, to some extent, mitigating the impact of converging AI and cyberthreats that target vulnerable subgroups and essential humanitarian services, as well as critical industrial, physical and information infrastructures** (Threat to the right to life, liberty and security of person, UDHR/Article 3).

▪ Second, given its increasing reliance on data-driven and predictive analysis, the field of peacekeeping needs to develop **the internal capacity and resource to prevent and mitigate potential threats to the integrity of its own datasets and analytical processes**. As explained at the end of section 3, effective mechanisms should secure the diverse digital repositories (from ad-hoc reporting systems in missions to more centralised efforts) that process sensitive data about situational awareness and populations' routine activities. Without robust (AI and cybersecurity) safeguards, data-driven peacekeeping could lead to unintended harm and erode public trust. Far beyond violations of privacy, unintended harm includes collective data breaches with serious security implications, especially when data is gathered from

[61] A cryptographic checksum is a mathematical value that is assigned to a file and used to "test" the file at a later date to verify that the data contained in the file has not been maliciously changed. A cryptographic checksum is created by performing a complicated series of mathematical operations that translates the data in the file into a fixed string of digits called a hash value, which is then used as a checksum. Without knowing which cryptographic algorithm was used to create the hash value, it is highly unlikely that an unauthorized person would be able to change data without inadvertently changing the corresponding checksum.

[62] A digital watermark is a hidden signal embedded into an image such that tampering corrupts the signal and thus indicates a loss of integrity.

vulnerable populations. The UN has already been the target of offensive cyber-attacks,[63] and **strong rules are needed to determine who will have access to sensitive information and what security measures will be used to ensure the integrity of the data**.

▪ Third, while challenges are rising when it comes to harnessing foresight, new skills and ensuring data-security and accountability, peacekeeping actors have a unique opportunity to approach technological and data-governance, including AI- and cyber-security, from **a conflict-sensitive perspective**. There is an urgent need to strengthen a **theory of no-harm in the data and technological convergence space** and such effort would benefit from both, **a forward-leaning and operational understanding of how data permeates the socio-technical systems of conflict**.

Based on the above diagnosis, this section will start by presenting potential **tech-centric responses**, in particular the AI and cybersecurity techniques that could support peacekeeping operations to prevent, manage and mitigate converging AI and security threats. While some originate from academic and civil society, most tech-centric responses are highly dependent on the techniques and tools fully or partially provided by corporate platforms and private security contractors in weakly regulated supply chains. Corporations with global reach already use automated predictive algorithms and other remote "threat-hunting" techniques for securing cybernetworks, monitoring user behaviour, and forecasting business risks, instability and armed violence.[64]

For the UN, increasing dependence on private sector technology raises important questions of responsibility and accountability and creates complex challenges to preserve political coherence and adherence to POC (and even human rights) considerations. To close the accountability gap, the field of peacekeeping will have to develop agreed and stress-tested methods to assess the ethical, security, and human rights implications of delegating some elements of POC mandates to new alliances of protection actors, including private firms in predictive analytics and threat-identification technologies. The UN Secretariat and related UN peacekeeping actors will also need critical **human-centric responses and skills**, including policy and foresight capacity, as well as cross-discipline and cross-sector collaborations, to mitigate technologies' dual-use potential and anticipate unforeseen normative and operational gaps in new hybrid conflict environments. Section 4 will close with a focus on these foresight-based and conflict-sensitive approaches to technological and data-governance.

INTEGRATING AI & CONVERGING TECHNOLOGIES IN POC MANDATE

What Skillsets and Technological Capacities?

What technological capacities would be needed to gain situational awareness in the context of increasing converging technological and security threats to civilian populations? What tools and techniques would be required to identify potential flashpoints of cyberconflict (what targets could a party be aiming to attack via AI and cyber means?) and to identify growth or decline in capability to wage AI, cyber and information operations (what capability does each party hold, are they becoming more advanced?)

Discrete forms of technological convergence could be harnessed to better manage complex information security risks to vulnerable populations and build resilience against rising forms of hybrid cyberattacks. For instance, AI's capacity for automating behavioural system analysis and anomaly detection already serves a role in cyber-defence, algorithms being able to detect abnormal and illicit behaviours across large computing networks and able to learn how to patch vulnerabilities against evolving cyberthreats. Several strategic functions within peacekeeping mandates could therefore be augmented by AI behavioural analysis and anomaly detection in cyber-defence.

[63] Marion Laurence, What are the Benefits and Pitfalls of 'Data-Driven' Peacekeeping?, Policy Brief, Center for International Policy Studies, University of Ottawa, December 2019.

[64] Companies like Palantir, Lockheed Martin, and GroundTruth have started exploring ways to turn situational awareness tools into better predictive matrices able to capture the interdependence of risk factors in conflict. Palantir, for instance, is using its expertise in predictive policing— algorithmic programs aimed at predicting the location and timing of crimes and violent attacks in cities—to better anticipate the strategies of terrorist groups in Syria.

Observation, Monitoring and Reporting (OMR): Peacekeeping operations could increasingly include technological capacity in AI and cybersecurity to monitor and, to some extent, protect a critical set of civilian services and systems related to industrial, physical and information infrastructures. This "cyberspace safe layer"[65] could include critical medical and humanitarian assistance; energy power grids; water management and supply chains; industrial safety, emergency and governance systems; as well as biometrics and electoral databases. Monitoring for threats to human security and the right to life (UDHR/Article 3) would be technically feasible, yet would still require the necessary consent from the host nation.[66] Such OMR function would not rely on or involve complex issues of attribution, but would consist in detecting threats to civilian life, warning essential prevention actors and, to the extent possible, providing patching or other technical countermeasures.

Similar to gaining situational awareness of a region, "cyber peacekeepers" would develop an acute understanding of what activities are normal or abnormal in network structure, harnessing AI and cybersecurity techniques to detect behavioural signals and anomalies upon the network they are monitoring. Across connected digital environments – from cloud computing services and edge-devices, to industrial control protocols and email systems – algorithms are able to ingest and analyse large swaths of data and interactions ("life patterns") within networks, and form an understanding of the normal behaviour of that environment. Using a self-learning approach, such algorithmic system constantly revises its understanding about what is normal based on evolving evidence.

- **Preventive Cyber-deployments**: In the context of preventive cyber-deployments (modelled after preventing peace deployments) or the establishment of a "cyber buffer zone,"[67] UN peacekeepers could rely on algorithmic defence systems to translate early warnings into early responses, neutralizing impending cyberattacks before they can harm the intended target. Private businesses and industries already rely on such AI-led prophylactic measures to protect their digital assets with automated threat-responses that are rapid and surgical. Human cybersecurity expertise is becoming more important at an overarching strategic level, but less needed to flag multitude of alerts and respond in real-time.

It would be important to assess whether this model of human-machine teaming in advanced cybersecurity fits the sensitive context of peacekeeping mandates (from OMR to establishing a "cyber buffer zone"). Figuring out this "human computation"[68] equation would be valuable in specific peacekeeping activities that aim at ensuring the cybersecurity of critical civilian services and infrastructures: what type of extensive cyber-monitoring can be automated through algorithmic defence systems? And how can this semi-automation support evolving forms of human reasoning and decision-making skills in (cyber-)peacekeeping? How do algorithmic defence systems support and work with strategic and operational expertise/knowledge of conflict? In peacekeeping context, assessing this "human computation" equation is critical, not only to optimize training and human skills acquisition, but also to anticipate unintended consequences, ensure accountability and conform with the principles of neutrality, impartiality and independence. For instance, if algorithmic defence systems can neutralize cyber-intrusion with an automated response, they would need to trace back and retain important evidence about the incident to fully serve conflict analysis, situational awareness and OMR activities.

- **Social Media Behavioural Analysis in Cognitive-Emotional Conflicts**: The combination of emotion analysis (affective computing), natural language processing (NLP), and speech/voice recognition technology allows for the mining of content within traditional and social media. These data streams comprising conversations, thoughts, and behaviours can help map local attitudes toward conflict and analyse emerging tensions, alliances, and divisions. They can also identify leaders and movements in fractured societies. Researchers at Princeton University demonstrated how machine-learning classifiers can learn to detect

[65] James, Joshua I & Breitinger, Frank. (2015). Digital Forensics and Cyber Crime: 7th International Conference, ICDF2C 2015, Seoul, South Korea, October 6-8, 2015. Revised Selected Papers. 10.1007/978-3-319-25512-5.

[66] Robinson, Michael & Jones, Kevin & Janicke, Helge & Maglaras, Leandros. (2018). An Introduction to Cyber Peacekeeping. Journal of Network and Computer Applications. 114. 10.1016/j.jnca.2018.04.010.

[67] Idem

[68] The field of human computation divides certain steps into computer and human activities and guides human performance. That means that a computer uses human abilities to solve specific problems and tasks, which are provided by a computer and cannot be solved by computers alone. This approach uses differences in abilities and alternative costs between humans and computer agents to achieve symbiotic human-computer interaction.

content that is part of coordinated influence operations based on human-interpretable features derived solely from content.[69] Widespread, automated influence campaign content leaves a specific signature in user-generated content that permits monitoring of campaigns over time and across different social media accounts. Through system behavioural analysis, algorithms may learn to detect emotional influencing and manipulation associated with rising ethnic and religious tensions leading to violence. Similar techniques could be applied to monitoring the proliferation of hate speech on social media in locations with high potential for conflict.

Calibrating human-machine teaming will be crucial when analysing cognitive-emotional conflicts involving ethnic subgroups, monitoring hate speech or conducting digital investigations and media forensics. Countering disinformation, for instance, has shown both the opportunities and limits of using anomaly detection and the need to pair algorithms with robust human expertise.[70] Situational understanding and knowledge of peacekeepers and conflict experts would be instrumental to the meta-data interpretation in cases where synthetic media and information manipulation is weaponized against certain ethnic subgroups.

- **Media Forensics and Digital Investigations**: The combination of AI behavioural detection and facial recognition could support human experts when they need to elucidate the actions of individuals and crowds during outbreaks of violence and human rights violations. For instance, Visual Forensics and Metadata Extraction (VFRAME) is a collection of image recognition software tools designed specifically for human rights investigations that rely on large data sets of visual media.[71] Another example is Forensics Architecture, a digital platform that conducts advanced spatial investigations into cases of atrocities, violent outbreaks, and human rights violations in urban warfare.[72]

Because they can be trained to detect anomalies, AI systems, when combined with precise image recognition, may increasingly play a role in virtual investigations in the context of information disorders in the run-up to elections, electoral assistance and ceasefire monitoring.[73] Engineers are also working on algorithms that can detect whether an image or video has been forged or tampered with.[74] If successful, algorithms for media forensics could uncover data manipulations, provide detailed information about the nature of these manipulations, and determine the overall integrity of visual media to facilitate decisions regarding the use of questionable images or video. Such algorithms will have to be resilient to adversarial attacks, which malicious actors could use to corrupt the anomaly detection process and blur investigations.

- **AI for Monitoring/Assessing Adversarial Cyber-Capacity:** The integration of AI's capacity for behavioural analysis and anomaly detection couldfocus on detecting changes in a state's cybersecurity dispositions (sudden augmentation of offensive cyber-capacity; or sudden reinforcement of cybersecurity measures to protect a certain type of critical infrastructure or industrial control systems); this may provide a way to forecast potential flashpoints of cyber conflicts and assess the sophistication of an actor's cybersecurity capability. Yet, in 21st century conflicts, understanding and anticipating state and non-state actors' capacity to cause civilian harms should go beyond a focus on cybersecurity and cyberwarfare, and integrate trends in technological convergence.

[69] Content-based features predict social media influence operations | Science Advances (sciencemag.org)

[70] AI-driven tools exist to counter disinformation by filtering fake news, reinforcing known facts, detecting nefarious content, eliminating trollbots, and verifying the authenticity of audio and video content. Yet, for now, technical limits and failures abound in these counter-disinformation systems. The main reason is that deep learning algorithms fail to understand contextual, linguistic, symbolic, and behavioural nuances of human online discourse. See: Douek E. 2018. "Facebook's Role in the Genocide in Myanmar: New Reporting Complicates the Narrative." Lawfare; 22 October. https://www.lawfareblog.com/facebooks-role-genocide-myanmar-new-reporting-complicates-narrative Also see: Wong JC. 2019. "'Overreacting to failure': Facebook's new Myanmar strategy baffles local activists." The Guardian; 7 February, https://www.theguardian.com/technology/2019/feb/07/facebook-myanmar-genocide-violence-hate-speech

[71] VFRAME is currently working with the Syrian Archive and the Yemeni Archive, organizations dedicated to documenting war crimes and human rights violations. VFRAME algorithms can process million-scale video collections, summarize scenes to reduce processing times, and detect objects such as illegal munitions. See VFRAME, "Computer Vision Tools for Human Rights Researchers," n.d., https://vframe.io

[72] Within virtual reality environments, Forensics Architecture locates, synthetizes, and analyzes pictures, videos, audio files, and testimonies from violence survivors to reconstruct and analyze conflict events. See Forensics Architecture, homepage, n.d., https://forensic-architecture.org

[73] "Election Interference to Be Sniffed Out by Early-Alert System," BBC News, 17 July 2018, https://www.bbc.com/news/technology-44820416

[74] Robert Bolles et al., "Spotting Audio-Visual Inconsistencies (SAVI) in Manipulated Video," University of Amsterdam, 2018, https://openaccess.thecvf.com/content_cvpr_2017_workshops/w28/papers/Mensink_Spotting_Audio-Visual_Inconsistencies_CVPR_2017_paper.pdf

Advanced technological capacity of state and non-state actors alike will become both, strategic to future conflicts and more difficult to assess and monitor. The decentralised supply chains through which converging technologies and related knowledge can proliferate and be outsourced as a paid service, could lead to unprecedented diffusion of power and dual-use potential in conflicts. Therefore, non-proliferation architectures have to, not only monitor procurement/transfer of technologies and equipment (for instance, surveillance technologies in cities across African countries partnering in China's Belt and Road Initiative), but also monitor intangible transfer of knowledge and technological know-how (including open and illicit platforms for sharing algorithms, training datasets, services and mentorship). Such non-proliferation challenges should become better integrated with peacekeeping mandates, including OMR activities.

How to Secure the Skillsets?

To be, either designed in house, or operated with due diligence, most of the techniques described above would require a specialised set of skills in AI (machine-learning, deep-learning and affective computing), data-analytics, automation and cybersecurity that are in high demand globally. As well noted by Di Razza and Mamiya, "DPO will need to consider which skills should be upgraded and professionalized (or outsourced) within the civilian component, which should be brought in from the military, and which can be an amalgam of both."[75] Potential sources of peacekeepers with this desirable skillset could include cyber military units from UN Member states, but also cybersecurity teams from private industry and academic talents.

In her 2020 IPI report, Martin-Brûlé emphasizes a set of professionalisation, training and retention challenges[76] in the field of peacekeeping intelligence: "refine the criteria for recruiting civilian and uniformed personnel with intelligence expertise" (recommendation 3); "improve retention of peacekeeping intelligence personnel and encourage member states to agree to longer-term deployments" (recommendation 4); "tailor peacekeeping intelligence training to the needs of missions while clarifying a standard set of UN norms" (recommendation 5). Her recommendations would apply to the field of AI and converging technologies with remaining questions about how the relevant expertise can be secured in the numbers required, and at a price that is within an operation's budget.[77] **There might be ways to organize specialized fellowship mechanisms or schemes for lending/providing particular skillsets in a sustainable manner, so that engineers in the private sector and academia can contribute to professionalisation of civilian staff in peacekeeping. Collaborations with AI and human rights labs in trusted partnerships with private companies could be envisioned as a starting point.**

Beyond this serious professionalisation challenge, two other policy and strategic issues pertain to the opportunities and risks for UN peacekeeping operations to increasingly rely on 1) "remote protection" and 2) procurement of private sector technology.

"Remote Protection"

Integrating AI and cybersecurity technological capabilities and related skillsets could support remote management of POC mandates by automating some elements of situational analysis and certain methods of monitoring populations' needs and routine activities. Yet, automated remote management may give peacekeeping operations a "false sense of informed decision-making"[78] and prevent them from assessing whether algorithmic monitoring is performing with accurate predictive value. There are significant ethical considerations about the potential harm caused by technical problems and failures in predictive value.[79] The limits to using AI and data capture technologies for predictive analysis of violent outbreaks and conflicts are significant: the lack of accurate, up-to-date, and representative data sets; the quality of data curation and algorithmic training; cognitive, gender, racial, historical, or economic biases; and a dearth of theoretical

[75] Di Razza and Mamiya, 2020, p. 10.

[76] Martin-Brûlé, 2020, p. 21.

[77] Robinson, Michael & Jones, Kevin & Janicke, Helge & Maglaras, Leandros. (2018). An Introduction to Cyber Peacekeeping. Journal of Network and Computer Applications. 114. 10.1016/j.jnca.2018.04.010.

[78] UNDPPA, "E-Analytics Guide," p. 18.

[79] Samuel Bazzi et al., "The Promise and Pitfalls of Conflict Prediction: Evidence From Colombia and Indonesia," NBER Working Paper, no. 25980 (June 2019).

and statistical knowledge about conflicts.[80] Conflict prevention actors must understand the computational techniques on which they rely and the data sets in use, particularly how data is collected and the biases those data sets may represent. When monitoring violence, human rights violations, or hate speech, it is crucial to measure the limitations of AI's predictive value and the incidence of false positives (violence is predicted but does not happen). The problem of "automation bias"—humans tend to stop questioning suggestions from automated decision-making systems and ignore contradictory information[81]—significantly raises the stakes for the use of AI in sensitive conflict analysis and prevention operations.

Building on local networks of peacekeepers with combined expertise in technology, intelligence and conflict-analysis will be crucial to avoid the risk of automation biases and predictive failures in remote management of POC mandates. Preserving the legitimacy of on-the-ground operations and securing the trust of vulnerable populations will also require community-based expertise. Localized staff should be able to verify the integrity of collected information and develop reliable analysis of how AI and converging technologies permeate systemic forms of political and social oppression. This is a follow-up reflection to crucial points made by Di Razza and Mamiya in their assessment of the skills and capacities needed on the ground to strengthen POC, namely: the need for local networks with community-based expertise to 'help missions avoiding finding themselves in support of predatory, illegitimate government actors being rejected by "civilians."'[82]

Private Sector Technology

Most successful, stress-tested and scaled-up technological capabilities in AI behavioural system analysis, anomaly detection and cybersecurity are the intellectual property of private companies. For UN peacekeeping operations, an increasing dependence on these advanced security technologies procured by private sector actors could raise issues of digital sovereignty, compromising consent of a host nation or undermining principles of neutrality, impartiality and independence crucial to UN mandates. In certain conflict-affected regions, such principles might be compromised if peacekeeping operations are using the capabilities of a private sector platform emblematic of a tech-power country.

Another area of caution is the complexity of dual-use technology supply chains. First, ensuring supply chain security would require UN actors to have or hire technical auditing capacity. For instance, procurement of algorithmic cyber-defence systems should rely on robust, independent audit and verification mechanisms to ensure the security of the supply chain (no backdoors or anomalies should be part of hardware and software products procured to UN peacekeeping). Second, the supply chains of converging and dual-use technologies are made of hybrid and unconventional actors, blurring traditional boundaries between civilian and military domains, between offensive and defensive purposes. The well-known precedent of Project Maven illustrates how large digital tech platforms increasingly consider selling civilian technologies to military organisations. Importantly, most AI technologies could play a powerful role in increasing the functionality of a wide-spectrum of dual-use applications that could indirectly serve, or be repurposed to serve, hostile intent of both, state and non-state actors alike. This spectrum is relevant to conflict situations and spans from intelligence, surveillance and reconnaissance (ISR) techniques (including image classification, language and speech recognition, GIS, predictive analytics, anomaly detection) to target tracking and recognition (TTR) and cyberspace operations. Private security contractors that perform military surveillance are expanding their expertise and tools in conflict analysis and threat-forecasting.[83]

Corporate normative and due-diligence frameworks currently exhibit serious shortcomings. A 2020 civil society report well explains how "most business and human rights initiatives and ethical standards fail to address many issues specific to rapidly changing technologies and their impact on human rights and conflict. They also fail to take into account how companies perceive, react to, and operationalise these norms at

[80] Weisi Guo, Kristian Gleditsch, and Alan Wilson, "Retool AI to Forecast and Limit Wars," Nature, no. 562 (October 2018), pp. 331–333, https://media.nature.com/original/magazine-assets/d41586-018-07026-4/d41586-018-07026-4.pdf

[81] M. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," American Institute of Aeronautics and Astronautics, n.d.,https://web.archive.org/web/20141101113133/http:/web.mit.edu/aeroastro/labs/halab/papers/CummingsAIAAbias.pdf

[82] Di Razza and Mamiya, 2020, p. 11.

[83] William Hartung, "Should Arms Makers Be Held Responsible for How Their Weapons Are Used?" Forbes, 9 September 2019, https://www.forbes.com/sites/williamhartung/2019/09/09/should-arms-makers-be-held-responsible-for-how-their-weapons-are-used/

scale."[84] The role that AI and converging technologies can play in conflict situations has become a rising concern for regulators and civil society.

Given the above challenges related to dual-use and proliferation, it is important for the field of peacekeeping to drastically upgrade its expertise to understand and foresee how converging technologies can impact the evolving conflict landscape.

Technological convergence will impact the evolution of POC mandates, and require the field of peacekeeping to develop robust human skills and expertise in: 1) monitoring the proliferation of dual-use technologies in conflict setting through decentralised supply chains (including through grey zone operations and related actors, economic development partnerships, and legacy systems from humanitarian assistance); 2) understanding how the integration of AI in offensive cyber- and information operations will threaten different aspect of civilian security; and 3) harnessing discrete technological defence capacities (such as AI systems for cyberthreat-monitoring and media forensics) to support OMR and prevention activities. **In brief, AI techniques and systems can only support but not replace substantial human skills and collective effort in sense-making and intelligence analysis, technological foresight and governance.**

POLICY AND FORESIGHT CAPACITY AT STRATEGIC AND OPERATIONAL LEVELS

Section 4 closes with overarching strategies that could serve to inform the future of peacekeeping operations. These strategies are purposedly "cross-boundary," based on "outside-the-box" thinking, and supporting but not limited to technological innovations. In the near- and longer-term, DPO and the Secretariat should plan for augmenting current peacekeeping expertise with new human-centric skills and responses to address the rapid changes and growing uncertainties posed by converging security threats in conflicts.

To Operationalize a Theory of No-Harm

There is an urgent need to strengthen a theory of no-harm in the data and technological convergence space and such effort would benefit from a **conflict-sensitive, operational understanding of how data permeates the socio-technical systems of conflict**. The ultimate rationale for having a theory of no-harm is to prevent the deployment of AI and other dual-use technologies in contexts where there are inadequate safeguards to protect human rights and insufficient mechanisms to ensure accountability. It would consist of devising policy and normative methods, for instance, critical data incident management mechanisms, to prevent civilian and human rights harms. These mechanisms would need to monitor all phases of technological design, development, and deployment, including a special focus on the life cycle of sensitive population data (data collection, retention, processing, and sharing). Dual-use and data-capture technologies in peacekeeping should only be deployed when their compliance with human rights can be demonstrated. This is where normative foresight and human rights impact assessment can play a role.

As a methodology, foresight can play a normative role and support a theory of no-harm by helping peacekeeping actors envision a range of scenarios on how to manage the tension between (1) improving predictive situational awareness analysis and (2) preventing or minimizing civilian and human rights harms generated by potential misuses of AI and data-capture. As an opportunity strategy, foresight methodologies can help peacekeepers and experts on the ground leverage ethical and normative solutions. As a form of interdependent risk management, these methods can help provide feedback loops to prevent or mitigate security, ethical, and governance failures across humanitarian systems and assistance sectors. Normative foresight efforts should imperatively include cooperation with civil society organisations in conflict-prone and conflict-affected areas as they have been at the forefront of the analysis and reporting of how behavioural engineering and information-manipulation may lead to collective data harms, other forms of civilian harms and human rights violations.

[84] JustPeace Labs, Technology in Conflict: Conflict Sensitivity for the Tech Industry, 2020. p. 2

Operationalising a theory of no-harm could start with the following multistakeholder efforts:

- Sharing due diligence and normative guidance and building policy capacity across the technology, policymaking, civil society, humanitarian, and peace-building sectors. In recent years, new cross-sectors and interdisciplinary partnerships, such as ICRC's DigitHarium, the GIFCT and the CyberPeace and Biometrics institutes, have allowed UN violence and conflict prevention actors, policymakers, and technology companies to engage on normative guidance, early-warning, and accountability mechanisms.

- Devising a common understanding of converging security risks in partnership with civil society and private sector actors to ensure coherence across those efforts and address knowledge gaps.

- Relying on contractual, technical, and organisational mechanisms to ensure that sensitive dual-use technologies remain in the hands of strategic actors in the humanitarian and peacekeeping sectors and do not spread to organizations that fall outside the scope of due diligence.

- Designing effective mechanisms and safeguards to secure population data sets. There is an urgent imperative to determine the right balance between proportionality in data collection, diligent sharing policy, and effective mechanisms for securing population data sets.

- Developing ongoing strategic foresight mechanisms, such as scenario-based and human-centric analysis, to anticipate accidental or purposeful misuse of dual-use technologies and their impacts on vulnerable populations. These mechanisms could forecast less predictable outcomes, such as technology being stolen or reverse engineered, or general purpose or civilian technologies being misused by states or violent nonstate actors.

Interesting discussions around the need for a "digital humanitarian space" have begun to tailor the procedures that govern digital and technological partnerships with private companies to the specific requirement (privileges and immunities) of the humanitarian and peacekeeping sectors. As well explained by Massimo Marelli, "what is essential is both, 1) wider political will on the part of external stakeholders to guarantee the protection of a digital humanitarian space, and 2) the awareness, knowledge, focus and determination of internal stakeholders to genuinely preserve the independence, impartiality and neutrality of international humanitarian organisations in cyberspace."[85]

Collective Sense-Making and Technical Foresight

For peacekeeping operations, what could matter substantially in the near-future is to develop an acute and agile **capacity for anticipatory strategic planning and sense-making** rather than focusing primarily on acquiring, centralising and retaining large collections of sensitive situational and population data. Collective foresight and future thinking methods could help peacekeeping staff being more adept at **making sense of the interdependence of signals and trends** that transpire from data flows and being able to **operationalise that knowledge on the ground**. There might be opportunities to harness broad national, academic and corporate expertise in foresight, systems thinking and new technologies and their applications in peacekeeping cooperation. Such cross-sector and cross-discipline cooperation could provide a framework, insights and a baseline intelligence to direct the efforts of a broader community of technologists from private and public domains, and to sharpen, elaborate, and "stress test" new peacekeeping strategies and innovations in increasingly volatile and unpredictable conflict environments.

- **Foresight Across-Technological and Security Sectors**: For instance, UN peacekeeping staff in collaboration with technologists, civil society actors, and policymakers could conduct **combined foresight analyses across technological and security domains** to anticipate and better understand emerging threats that could harness data-manipulation and target civilian populations, knowledge systems and critical infrastructures. This paper has shown how experts and stakeholders in data-driven fields – such as biomedicine and genomics – face rising AI and cyberthreats that target civilian security and public trust. They face growing concerns and challenges to ensure not only the confidentiality, but also the availability and integrity of datasets about vulnerable populations. Often, they work within relatively fragile information infrastructure

[85] Massimo, Marelli, Hacking humanitarians: Defining the cyber perimeter and developing a cyber security strategy for international humanitarian organizations in digital transformation (icrc.org), International Review of the Red Cross (2020), 102 (913), 367–387. Digital technologies and war, p. 387.

and with outdated cybersecurity frameworks. **It might be interesting to create opportunities for thinking about threat-identification and mitigation strategies within core groups of experts and technologists across cybersecurity, genomics security and data-driven (or cyber-) peacekeeping.**

These cross-discipline and multistakeholder groups could focus on the following foresight priorities:

• Learning from and with the AI, cyber-security (and genomics security) community. At the confluence of AI and cyberthreats, there is a need to explore and potentially implement red teaming, formal data-authentication and verification protocols, and responsible disclosure of AI/cyber vulnerabilities.

• Exploring different data-restriction, data-sharing and openness models. As the potential for adversarial data-manipulation rises, this paper highlights the need to consider norms, data- and algorithms-sharing regimes that favour safety and security, and other lessons from dual-use technologies.

• Promoting a culture of responsibility. AI and cybersecurity researchers, in collaboration with UN peacekeepers and conflict prevention experts are in a unique position to shape the security landscape of AI- and tech-enabled conflicts. Recent reflections have highlighted the importance of conflict-sensitive standards, norms, and expectations.

• Developing technological and policy solutions. High-level areas for further research include data-protection (confidentiality, availability, but also integrity), and coordinated use of AI and converging technologies for public-good security and POC mandates.

Models for Forecasting Public-Private Partnerships

Di Razza and Mamiya have emphasized the importance of "forecasting public-private partnerships" and "begin exploring serious policy proposals and new coordination frameworks for new types of protection actors, including how they are vetted."[86] In this regard, two organisations that aim at preserving civilian security in an era of technological convergence can provide useful models for peacekeeping as they have been able to bridge public and private efforts while learning from futures deep dives, and stress-testing planned strategies against changing operating environments.

-The US Cybersecurity and Infrastructure Security Agency (CISA): has learned to develop high-level preparedness and anticipatory intelligence across infrastructures, from electoral, industrial to health sectors. CISA relies on a unified and coordinated framework to "strengthen the security, resilience, and workforce of the cyber ecosystem that protect critical services and American way of life."[87] CISA engages in partnerships with private sector actors, communities, and government at every level to help make critical infrastructure more resilient to cyber and physical threats. In addition, CISA also focuses on ensuring that emergency preparedness communities can seamlessly and securely communicate during emergency operations to keep civilian populations safe, secure, and resilient.

-The International Gene-Synthesis Consortium (IGSC): shows that it is possible to perform security screening for converging technologies as a global, third-party authority, independent from national sovereignty concerns. IGSC has built strategic incentives and capacity for the private sector and academia to participate in policy and non-proliferation screening related to the biosecurity sector. IGSC has created a common global screening platform to help prevent the accidental or intentional misuse of DNA synthesis technologies. The advantages of such a platform include: 1) allowing a common way to access and update screening algorithms as new threat factors are identified; 2) ensuring economies of scale, making it affordable for existing DNA synthesis companies to co-develop and maintain an updated screening system; and 3) and finally, reducing barriers to entry for new synthesis companies as they can take advantage of a universal screening capacity instead of developing their own proceeding without adequate know-how. **Such global, independent, non-proliferation effort, which is able to forecast converging threats, develop related security screening, and work with public and private sector actors, could offer a model to forecasting public-private partnerships at the confluence of converging technologies and peacekeeping.**

[86] Di Razza and Mamiya, 2020, p. 10.

[87] https://www.cisa.gov/cybersecurity

Finally, increasing efforts should focus on embracing diverse and community-based expertise, as well as the next-generation of local innovators, in the design, development, and testing of AI and converging technologies for peacekeeping to ensure a broad range of perspectives and understanding of potentially sensitive use cases.

▪ **Community-based and Cross-Humanitarian Foresight** ("Nothing about us, without us"): When opportune (outside of extremely sensitive or classified contexts), some elements of UN staff (within peacekeeping and conflict prevention's innovation cells) could partner with UN Innovation Labs and other "humanitarian innovators" (ICRC's DigitHarium, OCHA Centre for Humanitarian Data, the CyberPeace Institute, etc.) to conduct inclusive foresight exercises in collaboration with local communities, for instance, democratized innovation ecosystems. Such a bottom-up approach would open the foresight process to local teams of young technologists and innovators from start-ups and grassroots open innovation labs across conflict-prone environments.[88] It could also support civilian populations, in particular the next-generation, by enabling education and social trust, resilience and empowerment.

Author: Eleonore Pauwels is an international expert in the security, societal and governance implications generated by the convergence of artificial intelligence with other dual-use technologies, including cybersecurity, genomics and genome-editing. Pauwels provides expertise to the World Bank, the United Nations, private sector actors and governments officials. She also serves as Senior Fellow for the Global Center on Cooperative Security in New York. Her most recent research focuses on AI-cyberthreats prevention, the changing nature of conflict, tech and bio-sovereignty, and foresight and global security. Email: eleonore@eleonorepauwels.com

[88] Interesting and trusted innovations might come from community-based expertise such as the Hala Systems. Hala Systems' interoperable platforms work together to warn civilians and predict where warplanes take off, where they will likely hit, where the danger areas are, and whom the planes belong to. In order to effectively warn citizens of impending airstrikes, the team behind Hala Systems needed to create a human network comprised of trusted contacts, recruited teachers, engineers, and even farmers as potential plane spotters. The team supplemented information from the human network with acoustic data, collected from remote sensors hidden in treetops and tall buildings, that helped determine speeds and aircraft models.

TECH CONVERGENCE	AI FUNCTION	AI-CYBER THREATS IN CONFLICTS	CIVILIAN HARMS	ATTACKS TARGETED AT PEACEKEEPING	TECH-CENTRIC RESPONSES	HUMAN-CENTRIC RESPONSES
CONVERGENCE OF POPULATION DATASETS (“The Internet of Bodies and Minds”)	**AI Malware** capable of: data exfiltration, data manipulation, synthetic data injection **Algorithmic** Predictive & Behavioural Analysis, Synthetic Media, Automated Targeting and Profiling	**Collective Data Harms**: -Population Data Theft -Population Data Manipulation -New forms of Cybercrime Targeting Multimodal Identities **Cognitive-Emotional Conflicts**: -Information Disorders -Proliferation of Hate Speech -Election Interference -Social Unrest -Violent Ethnic Conflict	Behavioural Surveillance & Engineering Physical Harms or Targeted Attacks Violations of Data Privacy and Political Agency Automated Ethnic Profiling, Discrimination Closing of Virtual Civic Space	Surveillance, Harassment or Harm to Local Informants Precision Biometrics Attacks Social and Emotional Engineering Attacks Disinformation & Synthetic Media Targeting Peacekeepers ⬇ **EROSION OF PUBLIC TRUST**	AI Behavioural System Analysis & Anomaly Detection For Observation, Monitoring and Reporting Preventive Cyber-deployments Social Media Behavioural Analysis in Cognitive-Emotional Conflicts Media Forensics and Digital Investigations	Community-based and Cross-Humanitarian Foresight Combined Expert Foresight Analyses across Technological and Security Domains Operationalising Theory of No-Harm Data-Integrity Incident Management POC-focused Policy Planning
INTERDEPENDENCE OF KNOWLEDGE SYSTEMS	**AI Malware** capable of manipulating integrity of data & algorithmic models	Sabotage of Critical Knowledge/Research Infrastructure Sabotage of Security and Governance Systems	Dysfunction of Crucial Data-Analytics Systems (Dysfunction in POC) ⬇ **EROSION OF PUBLIC TRUST**	Adversarial Attacks on Peacekeeping Datasets, Intelligence & Situational Awareness Analysis ⬇ **DYSFUNCTIONAL HARM, EROSION OF ORGANISATIONAL TRUST**	AI Behavioural System Analysis & Anomaly Detection, Preventive Cyber-deployments AI for Monitoring/Assessing Adversarial Cyber-Capacity	Combined Expert Foresight Analyses across Technological and Security Domains Data-Integrity Incident Management POC-focused Policy Planning
AUTOMATION OF INFRASTRUCTURE AND INDUSTRIAL CONTROL SYSTEMS	**AI Malware** capable of manipulating integrity of automated protocols	Sabotage and Weaponization of Manufacturing Platforms, Supply Chains, and Industrial Control Systems	Lack of Essential Services, Physical Harm and Economic Cost (Dysfunction in POC & humanitarian assistance) ⬇ **EROSION OF PUBLIC TRUST**		AI Behavioural System Analysis & Anomaly Detection, Preventive Cyber-deployments AI for Monitoring/Assessing Adversarial Cyber-Capacity	Combined Expert Foresight Analyses across Technological and Security Domains Data-Integrity Incident Management POC-focused Policy Planning

SECTION 2: FUNCTIONAL MATRIX – HOW CAN PARTIES TO CONCLICT MISUSE CONVERGING TECHNOLOGIES? AI-CYBER THREATS IN CONFLICT