GLOBAL CENTER
ON COOPERATIVE SECURITY

POLICY BRIEF

Artificial Intelligence and Data Capture Technologies in Violence and Conflict Prevention

Opportunities and Challenges for the International Community

By Eleonore Pauwels
September 2020

The field of violence and conflict prevention is about to face an upheaval as it confronts renewed questions about its capacity to analyze data, mitigate risks, and exert normative leadership in an era of converging security threats. Threats to human rights and security triggered by artificial intelligence (AI) and data capture technologies will require peacebuilding and violence prevention actors to bridge the gap between early warning and response and anticipate new challenges.

In a time of national emergencies and a global public health crisis, violence and conflict prevention actors will have increasing access to cutting-edge technologies such as predictive and automated behavioral analysis. Corporations use automated predictive algorithms for securing cybernetworks, monitoring user behavior, and forecasting instability and armed violence. Without adequate foresight, risk assessment, and normative leadership, governmental and international conflict prevention efforts may gradually rely on new, enhanced forms of behavioral surveillance driven by technologies fully or partially made by private sector actors in weakly regulated supply chains.

The potential of these converging technologies to control and influence human behavior has direct implications for UN violence and conflict prevention efforts, as well as its human rights agenda, including efforts to counter terrorism and prevent violent extremism. The United Nations is not set up to address the spread of AI and data capture technologies in the hands of corporations and violent nonstate actors, nor is it able to deter its members consistently and effectively from using such technologies for unlawful purposes. At the same time, there is a growing recognition that global natural hazards, such as pandemics or climate change, can impact the stability of many states and exacerbate conflicts and will require large-scale prevention, adaptation, and mitigation efforts.[1] The coronavirus pandemic, for instance, has stalled international, regional, and national conflict resolution efforts in regions where vulnerable populations are suffering the most. Converging security threats do not arise neatly within state boundaries; they are wielded globally and must be mitigated across states.

1 Adam Day and Jessica Caus, "Conflict Prevention in an Era of Climate Change: Adapting the UN to Climate Security Risks," UN University Centre for Policy Research, 2020, https://i.unu.edu/media/cpr.unu.edu/post/3856/UNUClimateSecurity.pdf.

INTRODUCTION

Our daily lives, composed of digital, geospatial, and biological data, have become fodder for behavioral surveillance. Our biometrics and biodata, movements and consumption patterns, conversations, and emotions are collected by AI and data capture technologies. This convergence provides a powerful analytical tool for predicting potential violence and can potentially assist in conflict prevention efforts. Relying on a diversity of data streams, AI programs can automate anomaly detection and behavioral analysis or find patterns and anomalies within the baseline data on the behavior of individuals and populations (box 1).

BOX 1. Artificial Intelligence and Data Capture Technologies for Behavioral Monitoring

Human demographic, biometric, and facial recognition. Algorithms can identify biometric data such as facial features, fingerprints, hand and ear lobe geometry, DNA, and voice samples by detecting and processing patterns and shapes specific to an individual (for example, segmenting and indexing someone's iris scan). Algorithms that perform biometric data identification and verification may lack optimal training data sets, can be prone to error, and can amplify existing bias within biometric measurements. Algorithms are also acquiring the ability to analyze visual data that constitute demographic and personal information, such as gender, race, and age, and to collect information on the outward appearance of individuals.

Human action recognition. Algorithms are able to interpret and understand the visual world, and natural language processing allows them to understand different languages. They can thus analyze and interpret human actions, from "recognizing simple human actions such as walking and running" to recognizing "realistic human activities involving multiple persons and objects."[a] The ultimate function of such algorithms is not just to interpret human actions, but also to predict them.

Crowd analysis. Artificial intelligence can help detect and qualify crowd behaviors, map social interactions or "grouping" in crowds, and flag anomalous or atypical behaviors. Although crowd analysis has great potential to support epidemiological surveillance, it raises the prospect of political and social surveillance of events such as protests.

Affect and behavior recognition. Affect recognition is a technique within affective computing, a field that aims to interpret individuals' emotional states by teaching computer-vision algorithms to analyze their facial expressions and voice modulation, eye movements and pupil dilatation, gait, and bodily responses. Although affect recognition lacks scientific validity, the technique is already used in academia and industry to devise applications spanning medical pain management, retail advertising, headhunting, student evaluation, and even predictive policing and criminal justice.[b]

a M.S. Roo, "Human Activity Prediction: Early Recognition of Ongoing Activities From Streaming Videos" (IEEE International Conference on Computer Vision, Barcelona, Spain, November 2011), p. 1, http://cvrc.ece.utexas.edu/mryoo/papers/iccv11_prediction_ryoo.pdf.

b Kate Crawford et al., "AI Now 2019 Report," AI Now Institute, 2019, https://ainowinstitute.org/AI_Now_2019_Report.pdf.

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. Furthermore, AI systems can learn to interpret and predict human actions, as well as classify behaviors and emotions as "normal," "abnormal," or "harmful." The convergence of AI and data capture technologies has powerful implications for the changing nature of conflicts and the global security landscape, including in relation to violent extremism and terrorism. These technologies also impact the methods and practices employed by the United Nations and other actors involved in violence and conflict prevention.

This brief analyzes how the dual-use potential of these technologies is already having an impact on violence and conflict prevention. It examines how these converging technologies can be positively harnessed and potentially misused in the field. The brief will explain how the new paradigm of predictive behavioral analysis and population data capture is increasingly presented as a solution to challenges in humanitarian action, conflict prevention, and peace and security. It will also explore ethical considerations and concerns regarding behavioral surveillance, state and commercial data-sharing practices, human rights violations, civilian security, and experimentation with complex technologies among vulnerable groups. It concludes with recommendations for ensuring a do-no-harm approach to deploying these technologies for violence and conflict prevention.

CONVERGING TECHNOLOGIES: A PARADIGM SHIFT FOR UN VIOLENCE AND CONFLICT PREVENTION EFFORTS?

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.[2] There are signs that AI and data capture technologies are already migrating to the violence and conflict prevention spheres.[3] One obvious temptation for member states and UN agencies is to use these technologies for behavioral monitoring and intelligence collection in different parts of the prevention agenda, for example, monitoring recurring patterns in epidemics, disasters, hate speech, terrorism, and violent outbreaks.

Conflict prevention is rapidly becoming an element of UN peacekeeper mandates in which technological and data governance will have powerful and unprecedented implications. UN peace-building and conflict prevention actors have developed comprehensive expertise in analyzing the systemic drivers of conflict, identifying likely threats to peace, and anticipating how these may spread if left unaddressed.[4] They deal with increasingly complex threat environments, including terrorism and violent extremism.[5] On the ground, they must seek to maintain impartiality, build trust within sensitive political relationships and local networks of informants, and develop a deep knowledge of the societies

2 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.

3 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.researchgate.net/publication/328256257_Retool_AI_to_forecast_and_limit_wars.

4 World Bank, "Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict," 2018, https://openknowledge.worldbank.org/handle/10986/28337.

5 See, for example, Arthur Boutellis and Naureen Chowdhury Fink, "Waging Peace: UN Peace Operations Confronting Terrorism and Violent Extremism," International Peace Institute, October 2016, https://www.ipinst.org/wp-content/uploads/2016/10/1610_Waging-Peace.pdf; Chris Bosley, "Violent Extremist Disengagement and Reconciliation: A Peacebuilding Approach," *Peaceworks*, no. 163 (July 2020), https://www.usip.org/sites/default/files/2020-07/20200729-pw_163-violent_extremist_disengagement_and_reconciliation_a_peacebuilding_approach-pw.pdf.

and traditions in each conflict setting.[6] Conflict prevention relies on understanding the human context and structured, in-depth behavioral analysis of the affected populations.

UN experts need to integrate emerging technologies to digitize, share, and secure the information they collect from open sources, human informants, and data capture technologies.[7] They also need to monitor 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. These experts must scrutinize how online 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 field of conflict prevention faces significant challenges with regard to its capacity for data analysis, anticipation, and decision-making in times of crisis.[8] Conflict prevention actors have the opportunity to transform a field in which contextual expertise, human intelligence, trust, and interpersonal skills are key elements into a powerful discipline of predictive and automated behavioral analysis. The trove of digital behavioral insights collected by UN agencies might serve the interest of several actors. Governments and private sector industries may see an opportunity for commercial and geostrategic influence, as well as data and resource capture. For the United Nations, a major challenge will be to define the opportunities but also the limits and risks of gathering "behavioral intelligence" about vulnerable populations through the use of technologies designed by private sector actors in complex supply chains.

The increasing use of technology to collect 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 conflict prevention 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 and data capture technologies.[9] The United Nations is likely to be tempted to accelerate the adoption of automated algorithmic systems for managing the data flood and improve its capacity for predictive and diagnostic analysis.

TOWARD AUTOMATED AND PREDICTIVE BEHAVIORAL MONITORING

Conflict prevention is already a field in which UN agencies use data capture technologies and intelligence collection to map and understand recurrent conflict patterns and forecast potential crises. The expertise, data analytics tools, and monitoring technologies are obtained through UN-contracted private companies or in collaboration with private sector actors, which pose important sustainability and security challenges.[10] Intelligence collection[11] for conflict analysis includes different types of data sets and techniques (box 2).

Technological convergence presents the multilateral system with the opportunity to adopt two significant measures in modernizing conflict prevention: (1) targeted behavioral monitoring at the individual and population levels and (2) automated and predictive behavioral and situational analysis.

6 Sebastian von Einsiedel, "What Works in UN Resident Coordinator-Led Conflict Prevention: Lessons From the Field," UN University Centre for Policy Research, June 2018, https://i.unu.edu/media/cpr.unu.edu/attachment/2869/RC-Project-Book-Upd-29JUN18.pdf.

7 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.

8 Adam Day, "Can Data Save UN Peacekeeping?," *World Politics Review,* 21 February 2019, https://www.worldpoliticsreview.com/articles/27479/can-data-save-u-n-peacekeeping; "Use Science, Technology to Bolster World's Collective Security, Secretary-General Says at Round Table," 23 March 2018, https://www.un.org/press/en/2018/sgsm18953.doc.htm.

9 Abilova and Novosseloff, "Demystifying Intelligence in UN Peace Operations."

10 The UN Conflict, Mediation, and Digital Technologies Toolkit makes an important argument about UN dependence on the private sector. "The fact that many of the data analysis tools are privately owned poses challenges such as the sustainable use of the resource, as well as the secure management and storage of the information collected." See UNDPPA and Centre for Humanitarian Dialogue, "Digital Technologies and Mediation in Armed Conflict," March 2019, p. 18, https://peacemaker.un.org/sites/peacemaker.un.org/files/DigitalToolkitReport.pdf.

11 Ibid.; Abilova and Novosseloff, "Demystifying Intelligence in UN Peace Operations."

BOX 2. INTELLIGENCE COLLECTION FOR CONFLICT ANALYSIS TOOLS

Open source intelligence (OSINT) and social media intelligence (SOMINT): Data about population behavior and intent. Human actions leave real-time digital traces, such as transaction data from an individual's use of digital services, mobile phones, purchases, and web searches. In addition, operational metrics and other real-time data is collected by UN agencies, nongovernmental organizations, and aid organizations to monitor their projects and programs. Online information from social media networks and traditional news media in different languages is a rich source of open source intelligence that can be analyzed by algorithms and text-mining programs to determine social and cultural attitudes, intentions, and behavior. While often privately owned, data analytics tools are commonly used to produce tactical intelligence out of multiple source data sets.[a] Local informants and crowd-sourcing provide strategic information, which is actively produced by users through mobile phone–based surveys, hotlines, and user-generated maps.

Communication with local informants, conflict parties, and other stakeholders relies on an array of digital technologies, including social media platforms, online chat rooms (mainly Facebook and Twitter), instant messaging applications (WhatsApp, Signal, Telegram, Viber, Line), and audio- and video-conferencing tools.[b] The use of these tools in conflict prevention and mediation also generates large amounts of user data and metadata—data that is produced around a communication, but is not part of the communication's content.[c] User data and metadata form what is called a "shadow profile" or "digital phenotype" that can be used to predict individuals' behaviors and preferences and analyze their personal information. These profiles are at risk of being misused by malicious actors for behavioral surveillance and engineering.

Imagery and tech-driven intelligence: Visual and geospatial data about population subgroups and urban activity. In the last decade, UN peace operations started using a range of data-capture and monitoring technologies, including unarmed unmanned aerial vehicles, i.e., drones; geographic information systems; satellites; full-motion video; ground-based sensors; and infrared imagery of changing landscapes, traffic patterns, light emissions, and urban development. Drones, for instance, have become indispensable in collecting intelligence in vast operational areas, where they are the best tools for observing illicit movements. Digital maps also play a major role in helping conflict mediation teams monitor real-time developments on the ground, including violent incidents, areas of control, the position and movements of troops, and population movements.[d] The UN Digital Toolkit on the role of digital technologies in armed conflict mediation emphasizes that "[i]f the data is triangulated with social media analytics, it can provide advance information of potentially destabilizing events—a form of early-warning—or insights into the sources and promoters of violence, hate speech, misinformation, or disinformation."[e]

a UN Department of Political and Peacebuilding Affairs (UNDPPA) and Centre for Humanitarian Dialogue, "Digital Technologies and Mediation in Armed Conflict," March 2019, p. 18, https://peacemaker.un.org/sites/peacemaker.un.org/files/DigitalToolkitReport.pdf.
b Ibid., p. 20.
c Privacy International, "Doing No Harm in the Digital Era," 11 December 2018, https://privacyinternational.org/report/2509/humanitarian-metadata-problem-doing-no-harm-digital-era.
d UNDPPA and Centre for Humanitarian Dialogue, "Digital Technologies and Mediation in Armed Conflict.
e Ibid., p 18.

Targeted behavioral monitoring at the individual and population levels

The vast amount of digital information now generated by populations means that more of these routine behaviors can be understood through AI-led computing.[12] The analysis of routine activities could serve 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. Algorithms now have access to a high volume and rich variety of data sets collected by smart sensing technologies in mobile devices and within cities' infrastructures. For instance, technologies such as facial recognition, gait analysis, and mobile biometrics can already analyze the faces and bodies of individuals in moving crowds.[13]

Other sources include communications metadata and internet connection records, but also extend to location and activity tracking, financial transactions, and social media activity. Much of this information is not directly in the hands of the United Nations but can be acquired through partnerships with mobile service providers, data brokers, biometric and AI companies, and internet and social media platforms. Since 2016, UN Global Pulse, an initiative that attempts to bring real-time monitoring and prediction to development and aid programs, has been engaged in a partnership with Twitter to access data tools and repositories.[14] In 2019 the World Food Programme (WFP) launched a five-year engagement with Palantir to optimize food delivery to populations on a biometric registry.[15] 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.[16]

Automated and predictive behavioral and situational analysis

Algorithms are being used to analyze large data sets, as they can be trained to identify data patterns and make inferences. Quantitative and even qualitative analyses of populations and situations can be optimized and automated by algorithmic programs. Algorithms use anomaly detection to perform systemic behavioral analysis of crowds: they learn to find patterns and anomalies within baseline data about how crowds are supposed to behave.[17] Anomaly detection can be made to work with varying degrees of automation. A system might be programmed, for example, to look for certain behaviors that are prelabled as anomalous, such as running or moving erratically, breaking equipment, moving against traffic, or carrying a gun.[18] Algorithms are also trained to identify an array of violent behaviors such as fighting, punching, and stalking. A subfield of AI called deep learning relies on

12 M.H. Divyashree and C.S. Shivaraj, "Artificial Intelligence for Human Behavior Analysis," *International Research Journal of Engineering and Technology* 5, no. 6 (June 2018), https://www.irjet.net/archives/V5/i6/IRJET-V5I6353.pdf.

13 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.

14 Anoush Tatevossian, "Twitter and UN Global Pulse Announce Data Partnership," UN Global Pulse, 23 September 2016, https://www.unglobalpulse.org/2016/09/twitter-and-un-global-pulse-announce-data-partnership.

15 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.

16 World Bank, "United Nations, World Bank, and Humanitarian Organizations Launch Innovative Partnership to End Famine," 23 September 2018, https://www.worldbank.org/en/news/press-release/2018/09/23/united-nations-world-bank-humanitarian-organizations-launch-innovative-partnership-to-end-famine.

17 Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-World Anomaly Detection in Surveillance Videos," 2018, http://openaccess.thecvf.com/content_cvpr_2018/papers/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.pdf.

18 Shian-Ru Ke et al., "A Review on Video-Based Human Activity Recognition," *Computers* 2, no. 2 (June 2013): 88–131, https://www.researchgate.net/publication/285197344_A_Review_on_Video-Based_Human_Activity_Recognition. For "violence detection," see E. Bermejo et al., "Violence Detection in Video Using Computer Vision Techniques," 2011, https://www.cs.cmu.edu/~rahuls/pub/caip2011-rahuls.pdf. See also Sadegh Mohammadi et al., "Angry Crowds: Detecting Violent Events in Videos," European Conference on Computer Vision, 2016, https://link.springer.com/chapter/10.1007%2F978-3-319-46478-7_1.

algorithms that can learn from large-scale data analysis what "normal" or "abnormal" behavior should be.

Peter Asaro explains how algorithms rely on "the Model of Threat Approach"—they assume that "the world can be classified into clear categories, *i.e.* threats and non-threats, to provide accurate classifiers of what constitutes a threat, and predictors of the likelihood and risk from that threat."[19] This form of binary and somewhat reductionist analysis is a substantial departure from how experts approach and understand the inherent complexity of human actions, interactions, and decision-making in conflict situations. With the convergence of data capture technologies and algorithms, the global AI industry posits that significant amounts of raw information about human experience can be turned into actionable intelligence. By monitoring diverse data streams and replacing human "eyes and ears" on the ground, AI could become useful for threat monitoring and situational awareness.

UN conflict prevention actors will face two challenges when managing the integration of AI and behavioral data-monitoring technologies into their work. If technologies can be trained with data sets that accurately mirror conflict dynamics, they may provide increasing potential to detect and analyze, within granular behavioral data, early-warning signals and patterns, drivers, and signatures for violent outbreaks and conflicts. Yet, conflict prevention actors might be using automated predictive behavioral analysis in contexts that exhibit high levels of uncertainty, and they may neglect to measure the potential technical and predictive failures that come with noise and data scarcity. Factors that amplify such risks include low-quality data, difficulty in separating signals from noise, and difficulty simulating human decision-making.[20] Another rising concern is the risk of surveillance technologies being misused by authoritarian regimes or violent nonstate actors. Even if substantial technical challenges remain, the increased functionality generated by converging technologies could play a substantial role in future conflict analysis and mediation.

PREDICTING HATE SPEECH, ARMED VIOLENCE, AND URBAN WARFARE

In the longer term, violence and conflict prevention experts may be tempted to rely on predictive models that can learn to identify strategic data interference from a combination of diverse risk signals that comprise a mix of social and causal factors.

Social media behavioral analysis and hate speech monitoring

Converging technologies are changing the strategic communications environments in which conflicts play out. Both state and nonstate actors can feed their own narratives and mis- and disinformation to their constituents within and across borders. Often, these narratives and falsehoods deliberately aim to stoke ethnic, religious, or political conflict. With AI technologies that can synthetize data and generate forgeries, the craft of emotional manipulation becomes ever more powerful.

UN conflict prevention actors are developing real-time analysis of complex social behaviors involved in conflicts.[21] The combination of emotion analysis, natural language processing (NLP), and speech and voice recognition technology allows for the mining of content within traditional and social media. These data streams comprising conversations, thoughts, and behaviors can help map local attitudes toward conflict and analyze emerging tensions, alliances, and divisions. They can also identify leaders and movements in fractured societies. Through automated behavioral analysis, AI systems could detect behaviors associated with rising ethnic and religious tensions leading to violence. Interestingly, the perception of a "potential threat" defined by patterns of behaviors perpetuates the need for data capture and monitoring technologies that render these patterns visible.

19 Peter Asaro, "AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care," 2018, p. 3, https://peterasaro.org/writing/Asaro_PredictivePolicing.pdf.
20 Guo, Gleditsch, and Wilson, "Retool AI to Forecast and Limit Wars."
21 UNDPPA and Centre for Humanitarian Dialogue, "Digital Technologies and Mediation in Armed Conflict."

The UN Secretary-General's strategy and action plan on hate speech points to the need to use technology to understand the relationship between the misuse of social media for spreading hate speech and the factors that drive individuals to violence.[22] Combining speech recognition technologies with NLP, analysts could go beyond social media to monitor television and radio communications for signals of hate speech in real time. UN Global Pulse is in the process of developing algorithms to characterize, quantify, and automatically detect elements of hate speech from heterogeneous data sources. Global Pulse already has experience in mining radio content with machine learning and has applied its expertise to monitor the refugee crisis in Uganda.[23] The experience showed that analyzing speech content with NLP could bridge the information gap by providing early warning of an impending crisis. Similar techniques could be applied to monitoring the proliferation of hate speech on social media in locations with high potential for conflict. Private sector actors are also engaged in efforts to address abuses of their digital platforms and enable multi-stakeholder engagement around issues related to terrorist and violent extremist misuses of the internet, such as through the Global Internet Forum to Counter Terrorism (GIFCT).

Digital investigations and media forensics

The combination of AI behavioral detection and facial recognition could elucidate the actions of individuals and crowds during outbreaks of violence and human rights violations.[24] 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 election monitoring and conflict mediation.[25] Engineers are also working on algorithms that can detect whether an image or video has been forged or tampered with.[26] 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.

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.[27] 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.[28]

Predictive AI modeling for conflict

Major private sector actors in strategic intelligence have shown interest in using AI's capacity for anomaly detection and predictive behavioral analysis to anticipate and measure the risk of conducting business or building complex supply chains in countries prone to violence and conflict. Companies such as Palantir,

22 The Secretary-General's action plan on hate speech explains how relevant UN entities should be able to recognize, monitor, collect data on, and analyze hate speech trends, as well as adopt a common understanding of the root causes and drivers of hate speech. *UN Strategy and Plan of Action on Hate Speech*, May 2019, https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf.

23 John Quinn and Paula Hidalgo-Sanchis, "Using Machine Learning to Analyse Radio Content in Uganda," UN Global Pulse, September 2017, https://unglobalpulse.org/wp-content/uploads/2017/09/Radio-Analysis-Report_Preview.pdf.

24 G. Shivashree and S.G. Anuradha, "Crowd Analysis Using Computer Vision Techniques," *International Journal of Engineering Research in Computer Science and Engineering* 5, no. 4 (April 2018), https://www.technoarete.org/common_abstract/pdf/IJERCSE/v5/i4/Ext_42371.pdf.

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

26 Robert Bolles et al., "Spotting Audio-Visual Inconsistencies (SAVI) in Manipulated Video," University of Amsterdam, 2018, https://staff.fnwi.uva.nl/t.e.j.mensink/publications/bolles17cvprwmf.pdf.

27 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 (accessed 19 September 2020).

28 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.

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.[29] Palantir had been criticized for a lack of transparency when it ran an algorithmic scoring system for predicting crime in New Orleans without the consent of city council members and prior public debate.[30] The predictive technology is now being adapted and used by Israeli security services.[31]

From a violence prevention perspective, there is a crucial need to better understand and predict when and why people turn to violence. It is also important to appreciate the evolving nature of intrastate wars, including the complex human decision-making processes of armed groups and their strategies to recruit youth and blend into civilian environments. Yet, due diligence must be performed when interconnections and alleged portability of analytical tools and methods are uncovered, whether by predictive policing, military intelligence, or conflict prevention.[32] Another area of caution is the complexity of dual-use technology supply chains; corporations that supply weapons and perform military surveillance are also expanding their expertise and tools in conflict analysis and prevention.[33]

Humanitarian actors operate under the principles of humanity, neutrality, impartiality, and independence. UN peace-building actors will have to seriously weigh the methods, technologies, and alliances they want to utilize in this new paradigm of behavioral surveillance (fig. 1).

FIG 1. CONVERGING TECHNOLOGIES FOR VIOLENCE AND CONFLICT PREVENTION

AI METHODS/FUNCTIONS	VIOLENCE AND CONFLICT PREVENTION PURPOSE
Human demographics, biometrics, and facial recognition Algorithms can identify biometric data by detecting and processing patterns and shapes specific to an individual. Algorithms are acquiring the ability to analyze visual data that constitute demographic and personal information such as gender, race, and age and can collect information on the outward appearance of individuals.	**Example:** To identify targeted individuals in digital investigations; to identify which individuals are part of armed groups

29 Ali Winston, "Palantir Has Secretly Been Using New Orleans to Test Its Predictive Policing Technology," *Verge,* 27 February 2018, https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd.

30 Ibid.

31 Orr Hirschauge and Hagar Shezaf, "How Israel Jails Palestinians Because They Fit the 'Terrorist Profile,'" *Haaretz,* 31 May 2017, https://www.haaretz.com/israel-news/.premium.MAGAZINE-israel-jails-palestinians-who-fit-terrorist-profile-1.5477437.

32 Andrew Selbst et al., "Fairness and Abstraction in Sociotechnical Systems," Association for Computing Machinery, 2019, http://sorelle.friedler.net/papers/sts_fat2019.pdf.

33 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/#52bc191f1709.

FIG 1. CONVERGING TECHNOLOGIES FOR VIOLENCE AND CONFLICT PREVENTION (continued)

AI METHODS/FUNCTIONS	VIOLENCE AND CONFLICT PREVENTION PURPOSE
Human behavior and action recognition Algorithms can learn to analyze and interpret human actions. For instance, artificial intelligence (AI) systems might be programmed to look for certain behaviors that are predefined as anomalous, such as "running or moving erratically, loitering or moving against traffic, or dropping a bag or other items," and to "perform detection of various kinds of violent behaviors such as fighting, punching, and stalking." Increasingly, the function of algorithms is not just to interpret human actions, but also to predict them.	**Example:** To capture, analyze, and predict behaviors of armed groups or clashes between rebel groups
Crowd behavior and action analysis Algorithms learn to detect anomalous or atypical behaviors in crowds, perform "multiple people tracking," and understand "grouping in crowds"—finding social connections between individuals.	**Example:** To capture, analyze, and predict behaviors of large crowds during protests, breaking of ceasefires, migration, or population movements before impending attacks; to aim to predict where a group of enemy forces will move next
Cybercrime and cyberintrusion analysis AI systems will be able to detect illicit online cybercrime activities that could be used by violent nonstate actors. AI systems will soon be probing and defending cybernetworks with very high degrees of autonomy and with a level of speed and complexity that far surpasses human understanding. Given the rapid increase in the speed and volume of cyberattacks, the use of completely autonomous tools for cyberwarfare appears inevitable.	**Example:** To protect critical information and cybersecurity assets of the United Nations, including situational awareness and conflict analysis data sets, tools, and maps
Geospatial mapping for situational awareness analysis Remote sensing imagery from satellites or drones can give important insights into conditions on the ground, including in areas that are difficult to access.	**Example:** To assess operational risks and integrate incidents such as theft of livestock, as well as environmental conditions (from signs of food insecurity to signs of water scarcity and depletion)

Source: Global Center.

ETHICAL AND HUMAN RIGHTS CONSIDERATIONS

The field of conflict prevention increasingly relies on AI and data capture technologies to collect large amounts of behavioral and contextual information about populations living in fragile countries or those prone to violent outbreaks. Such troves of digital information are, however, at risk of being misused by states, corporations, and violent nonstate actors for precision surveillance.[34]

HUMAN RIGHTS IMPLICATIONS OF BEHAVIORAL SURVEILLANCE

The use of systemic, automated, and predictive behavioral analysis in violence and conflict prevention is likely to pose new and fundamental challenges for protecting human rights in fragile contexts. Although these technologies may present long-term gains, there are also ethical risks associated with an AI-driven strategy for violence and conflict prevention. UN agencies could potentially be collecting and managing growing amounts of sensitive data about vulnerable populations and adopting behavioral monitoring technologies partially or fully made by private sector actors in complex, weakly regulated supply chains.

In this era where AI combines with powerful data capture technologies such as biometrics and facial and emotion recognition, algorithmic surveillance amplifies "biopolitics," a series of interventions to regulate society's collective body.[35] There is a tendency to underestimate how converging technologies can be designed to anticipate and influence human behaviors 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, exposure to pervasive data security breaches, and new forms of censorship in the virtual civic space.

Accountability and Remedy. Any autonomous data capture and algorithmic system that affects life and death scenarios is a test of the norms of remedy and accountability. Audit studies have shown that AI systems can act in unpredictable ways.[36] If such a system were diverted from its initial purpose and misused to harm human rights, it would be difficult to ascertain responsibility. Responsibility might be shared between the designer of the system; the supplier of the system; a third party, such as a vendor; or the international organization sponsoring the use of AI in conflict prevention. Some experts believe that using AI systems will create an accountability gap, making it difficult if not impossible for injured parties to access a remedy or receive fair treatment by the justice system.[37] There is also considerable concern that AI technologies could weaken the international rule of law, for example, through the facilitation of extrajudicial actions such as lethal drone strikes during intelligence and surveillance activities.[38]

Right to Privacy. Converging technologies pose profound challenges to privacy because they can automate the detection of "anomalies" or "abnormal behavior" in bulk data routinely collected about individuals and crowds. For example, image and speech recognition algorithms can detect objects in blurry photographs or separate voices in crowded environments. The dual-use potential of these technologies is problematic because technologies built for lawful uses can easily be adapted to facilitate surveillance in violation of human rights principles.

Right to Self-Determination. By introducing new opportunities for authoritarian states or violent nonstate actors to control populations, AI threatens

34 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.

35 Michel Foucault, *The History of Sexuality: An Introduction,* vol. 1 (London: Vintage Books, 1976), pp. 138–139.

36 Roman Yampolskiy, "Unpredictability of AI," University of Louisville, 29 May 2019, https://arxiv.org/ftp/arxiv/papers/1905/1905.13053.pdf.

37 Rachel Courtland, "Bias Detectives: The Researchers Striving to Make Algorithms Fair," *Nature*, no. 558 (20 June 2018), https://www.nature.com/articles/d41586-018-05469-3.

38 ICRC, "Artificial Intelligence and Machine Learning in Armed Conflict."

political participation and freedom of movement. In the same way, AI-led surveillance technologies also pose a real threat to peaceful assembly and protest.

Right to Freedom of Expression. When information disorders threaten conflict and mediation efforts, UN actors might find themselves caught in a debate about free speech versus removal of content.[39] This debate would take place alongside the work already being done by law enforcement agencies, intelligence agencies, civil society organizations, and technology companies. Yet by developing AI capabilities to better understand, analyze, and detect signs of hate speech and incitement to violence during elections, UN mediators could also play a key role in protecting political participation in the digital age.

Nondiscrimination and Minority Rights. Current facial recognition algorithms are sometimes unable to discern the features of darker-skinned faces with the same rates of accuracy as they detect the geometry of lighter faces.[40] Minorities could be stigmatized and ostracized in new and powerful ways. It is worth considering how this issue or another unintended consequence could have corrosive implications for digital investigations and conflict monitoring efforts. Another major concern is the potential for automated ethnic profiling or "techno-racism."[41] 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.[42] In recent years, multiple investigations by human rights defenders have also unveiled how the Chinese government imposed facial recognition tracking and the collection of biometric data, including DNA samples and voice samples, on its Uighur population.[43] Chinese authorities now have in place a vast system of facial recognition algorithms that have been trained to associate certain skin tones and facial features with the Uighur ethnicity.[44] This type of profiling makes China a leader in applying AI to monitor subpopulations, with the potential to export a new type of automated racial surveillance.

CIVILIAN PROFILING IN CONFLICT

Conflict prevention and mediation actors use digital networks to analyze and understand insights about populations' attitudes, behaviors, and intentions. They harness social media networks, AI-powered platforms, private communication channels, and mobile phones to elicit and analyze information about peoples' attitudes in conflicts and during times of ethnic or socioeconomic tensions.

Sensitive data about individuals' locations, livelihoods, behavior, and intentions can potentially be subject to an array of powerful surveillance practices by malicious actors. Population subgroups could be targeted by states or violent nonstate actors for the information they share online about hate speech, violence, election fraud, the activities of armed groups, and impending attacks. The UN staff's reliance on social platforms and digital channels also creates abundant metadata that can be used to "predict people's behaviors, preferences, and other personal details (e.g. ethnicity, sexual

39 Chris Westfall, "The Free Speech Face-Off Between Facebook and Twitter: Are Warnings Justified?" *Forbes*, 30 May 2020, https://www.forbes.com/sites/chriswestfall/2020/05/30/free-speech-facebook-twitter-george-floyd-demonstrations-censorship/#66f54d726e90.

40 Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research,* no. 91, 2018, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

41 In her June 2020 report, the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia, and related intolerance analyzed 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).

42 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.

43 Maya Wang, "Eradicating Ideological Viruses: China's Campaign of Repression Against Xinjiang's Muslims," Human Rights Watch, 2018, https://www.hrw.org/report/2018/09/09/eradicating-ideological-viruses/chinas-campaign-repression-against-xinjiangs; Danny O'Brien, "Massive Database Leak Gives Us a Window Into China's Digital Surveillance State," Electronic Frontier Foundation, 1 March 2019, https://www.eff.org/fr/deeplinks/2019/03/massive-database-leak-gives-us-window-chinas-digital-surveillance-state.

44 Paul Mozur, "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority," *New York Times,* 14 April 2019, https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

orientation, political and religious affiliations)."[45] Such digital profiles can be exploited for surveillance, harassment, and behavioral control and can endanger the safety of informants and mediators.[46] For instance, 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.[47] In the near future, tracking informants could rely almost solely on the data captured from mobile communications or elements of their digital identity revealed through online interactions.

EMOTION WARS: MANIPULATING POPULATION BEHAVIORS

The use of predictive behavioral monitoring within conflict analysis and prevention efforts may also lead to more precise, actionable intelligence about the root causes of ethnic and socioeconomic tensions, including warnings about hate speech and incitements to violence. Yet, this type of analysis or intelligence could also inform disinformation campaigns and emotional manipulation techniques by factions in conflict, from ruling elites and political parties to terrorist groups.

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.[48] The violent anti-Islamic backlash that followed was then instrumentalized by ISIL for its recruiting strategies. Since then, digital and emotional manipulation techniques have become both democratized and refined. In August 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.[49] Algorithms that do not need to be trained on each new face target provide a powerful tool kit to create realistic video forgeries at scale and with minimal know-how.

It is crucial to analyze and understand how populations facing ethnic tensions or residing in conflict-prone regions might perceive threats of synthetic media in a different light than those living in the West. Citizens from South Africa expressed significant concerns over the use of deepfakes by ruling political parties to incite violence beyond generating disinformation.[50] Their concerns were heightened by the potential for deepfakes to spark mob violence in areas suffering from political or ethnic tensions. Similar incidents had already happened in the country when manipulated videos were spread to encourage xenophobic attacks targeting Nigerian businesses.[51]

Predictive behavioral and emotional analysis already enables hyperpersonalized political campaigns in which key demographics are manipulated to affect

45 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.

46 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.

47 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.

48 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.

49 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.

50 Corin Faife, "In Africa, Fear of State Violence Informs Deepfake Threat," Witness, 9 December 2019, https://blog.witness.org/2019/12/africa-fear-state-violence-informs-deepfake-threat.

51 Ibid.

voting behavior.[52] Yet, as witnessed in cybercrime, the combination of psychometric tools and emotional engineering using personal data sets can help craft attacks that are hardly recognizable as such.[53] In the near future, malicious actors will be able to rely on predictive behavioral 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. In countries where there are few accountability mechanisms for privacy and data protection, domestic political parties can exploit sensitive population data sets and social media networks for spreading targeted propaganda, hate speech, and mis- and disinformation. The deployment of AI-enabled forgery technology will drastically alter the relationship to evidence and truth across journalism, criminal justice, conflict investigations, political mediation, and diplomacy. The capacity of a range of actors to influence public opinion with misleading information could have powerful long-term implications for the role of the United Nations in maintaining peace and security.

AI EYES AND EARS ON THE GROUND: MAPPING DIGITAL BODIES

Beyond social media and open source intelligence, conflict prevention actors will likely collect growing amounts of behavioral information on populations using automated data capture technologies. Access to the large data sets captured through satellites and drones equipped with video surveillance can reveal the location, settlements, and movements of ethnic subgroups, minorities, and refugees. Such data is translated into dashboards and digital maps generated for improving situational awareness, locating outbreaks of violence, and identifying civilian populations and infrastructure in need of protection.[54]

These risks have implications for conflict prevention operations but are also heightened by the increasing permeability of critical data collected in humanitarian operations and conflict prevention situations. For instance, biometric data from Syrian refugees are systematically collected to create a form of "cross-border identity" in complicated displacement situations.[55] "[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."[56] A second example is the combination of remote satellite imagery and machine learning to localize and visualize refugee settlements on digital maps. Scholars and officials of the UN Institute for Training and Research's Operational Satellite Applications Programme stress that "with the emergence of high-resolution high frequency imagery that is easily accessible, it is imperative that we seriously consider the privacy implications and the potential unintended consequences of sharing or using satellite imagery."[57]

It is increasingly difficult to control who can and should access population data sets and the subsequent visual maps and actionable intelligence they produce. Data collection, analysis, and visualization require the involvement of an array of actors external to the United Nations, from civil society organizations to private sector actors and governments. Recent deconfliction attempts also revealed the threat of data misuse as discussed on 30 July 2019 during a briefing to the Security Council on the humanitarian situation in Syria. Within a deconfliction process,

52 Alex Hern, "Cambridge Analytica: How Did It Turn Clicks Into Votes?" *Guardian,* 6 May 2018, https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie.

53 Eleonore Pauwels, "The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI," UN University Research and Publications, 2 May 2019, https://cpr.unu.edu/the-new-geopolitics-of-converging-risks-the-un-and-prevention-in-the-era-of-ai.html.

54 UNDPPA, "E-Analytics Guide."

55 Madianou, "Biometric Assemblage."

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

57 John A. Quinn et al., "Humanitarian Applications of Machine Learning With Remote-Sensing Data: Review and Case Study in Refugee Settlement Mapping," *Philosophical Transactions of the Royal Society* 376, no. 2128 (13 September 2019): 13, https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2017.0363.

humanitarian agencies, mainly nongovernmental organizations, provide information to the UN Office for the Coordination of Humanitarian Affairs to identify static civilian locations or humanitarian movements. The United Nations then shares this information with the parties to the conflict with the hope that civilian infrastructure such as hospitals will be spared. Yet, in northwest Syria in 2019, six different attacks targeted deconflicted civilian locations and humanitarian movements. 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."[58]

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 conflict prevention actors progressively adopt the deployment of AI technologies for situational awareness, population monitoring, and digital investigations. Exported to the global South, algorithmic surveillance is already transforming cities into networks of sharp eyes, where biometrics and facial and emotion recognition are used to analyze human actions for predictive policing. In 2018, as a beneficiary of China's Belt and Road technological development program, the government of Zimbabwe signed a series of contracts to deploy networks of CCTV cameras connected with facial-recognition software across cities' infrastructure, from public transport (bus stations, railway, and airports) to health facilities and smart financial systems.[59] The Chinese company CloudWalk is deploying its 3D light facial software, which is more accurate at detecting the facial features of dark-skinned populations, to build a database of Zimbabwean citizens, with their faces likely matched with other biometrics.[60] Such a national identity system, paired with compulsory SIM card registration data, provides the architecture for precision population surveillance, from physical movements to online activities.

CIVILIAN DATA AND INFORMATION INFRASTRUCTURE SECURITY

As conflict analysis increasingly integrates AI and data capture technologies, UN agencies will face heightened cybersecurity risks. Conflict analysis already relies on AI predictive analytics, and this trend is likely to accelerate. 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 behavioral analysis also have the potential to modernize the field of conflict prevention.

Moving forward, conflict prevention actors will need to thoroughly secure the behavioral and contextual information they collect about populations. Protecting such sensitive data sets from digital manipulation and cyberexfiltration will be a complex challenge, but it is crucial to ensuring the security of civilians and critical information infrastructure. For instance, remote sensing imagery from satellites or drones can provide information about refugee settlement, the precise locations of populations, human rights violations, and structural damage from conflict. A cyberattack by a state or violent nonstate actor could potentially exfiltrate sensitive information for surgical offensives or strikes.

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

58 Mark Lowcock, "Briefing to the Security Council on the Humanitarian Situation in Syria," UN Office for the Coordination of Humanitarian Affairs, 30 July 2019, https://reliefweb.int/sites/reliefweb.int/files/resources/ERC_USG%20Mark%20Lowcock%20Statement%20to%20the%20SecCo%20on%20Syria-%2030July2019%20-%20as%20delivered.pdf.

59 Lynsey Chutel, "China Is Exporting Facial Recognition Software to Africa, Expanding Its Vast Database," Quartz Africa, 25 May 2018, https://qz.com/africa/1287675/china-is-exporting-facial-recognition-to-africa-ensuring-ai-dominance-through-diversity.

60 Samuel Woodhams, "How China Exports Repression to Africa," *Diplomat,* 23 February 2019, https://thediplomat.com/2019/02/how-china-exports-repression-to-africa.

weaknesses of Syrian opposition activists.[61] Today, AI malware can watch, track, and evaluate individuals' emotions, language, and behavior, 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 behavior 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.[62] In the near future, autonomous malware could tailor their offensive strategies to the human targets they need to manipulate, including individuals within networks of civil informants, mediators, and UN staff. In the context of humanitarian assistance, the vulnerability of biometric databases will remain a constant and long-term concern.

LOGIC OF EXPERIMENTATION AND TECHNICAL AND PREDICTIVE FAILURES

AI and data capture technologies may improve the ability to automate detailed conflict analyses from afar and remotely monitor the needs of vulnerable populations, particularly in the midst of a pandemic. Yet, automated remote management may give UN agencies a "false sense of informed decision-making"[63] 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.[64] 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 and statistical knowledge about conflicts.[65] 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[66]—significantly raises the stakes for the use of AI in sensitive conflict analysis and prevention operations.

TOWARD A THEORY OF HARM

There is an urgent need to devise a theory of harm in the AI space. The conflict prevention sector and, to some extent, the humanitarian sector have not yet fully developed and operationalized a common "theory of harm." Such a theory would consist of developing adequate methods to weigh the benefits of harnessing converging technologies, in particular automated and predictive behavioral monitoring, against the costs to civilian security and human rights. Within the UN conflict prevention platform, there are no agreed and stress-tested methods to assess the ethical, security, and human rights implications of delegating some elements of intelligence collection and conflict analysis to automated predictive technologies. A substantial accountability gap exists because there are no methods and cross-sector collaborations to anticipate unforeseen misuses and long-term impacts of AI and data capture technologies on vulnerable populations.

61 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.

62 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.

63 UNDPPA, "E-Analytics Guide," p. 18.

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

65 Guo, Gleditsch, and Wilson, "Retool AI to Forecast and Limit Wars."

66 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.

The ultimate rationale for having a theory of harm is to prevent the deployment of AI, biometrics, 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, human rights impact assessments, to weigh the benefits of harnessing dual-use technologies for violence and conflict prevention against the costs to civilian security and human rights. These assessments 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 technologies in violence and conflict prevention should only be deployed when their compliance with human rights can be demonstrated.

As a methodology, foresight can play a normative role and support a theory of harm by helping conflict prevention actors envision a range of scenarios on how to manage the tension between (1) improving predictive conflict analysis and bridging the warning-response gap and (2) preventing or minimizing civilian and human rights harms. As an opportunity strategy, foresight methodologies can help conflict prevention actors 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 systems and sectors. For instance, conflict prevention actors could work on a set of normative issues such as data privacy and security implications and accountability and inclusion gaps. Normative and inclusive foresight also should include civil society organizations who are at the forefront of the analysis and reporting of how behavioral surveillance may lead to human rights violations. One ultimate benefit of engaging local civil society in normative foresight is not to deploy AI and converging technologies for conflict prevention in cases where there are inadequate safeguards to protect human rights.

In 2018 and 2019, many companies produced AI ethics principles and due diligence statements.[67] These efforts had several shortcomings. First, most of these statements were conceived by organizations whose leaderships operate in the global North. Statements of principles and ethical priorities of the global South, as well as those of populations affected by violent conflict, are often absent from these normative AI maps. Second, there are obvious limits to self-regulation and corporate ethical principles. These principles need to be translated and turned into viable normative practices that can be overseen and tested for transparency and accountability. Third, technological convergence and its implications are rarely understood when private sector actors define technical and normative standards. Responses to these three shortcomings must be discussed by UN agencies, civil society, and private sector actors under the umbrella of international human rights and humanitarian law and the UN Guiding Principles of Business and Human Rights.

RECOMMENDATIONS

The field of violence and conflict prevention will soon face one of its most difficult strategic and ethical choices: whether it should fully embrace and integrate converging technologies designed for predictive analysis based on the behavioral data of mass populations. Without adequate foresight, risk assessment, and normative leadership, prevention actors may come to rely on new and enhanced forms of behavioral surveillance driven by technologies fully or partially made by private sector corporations. Such a transformational shift raises significant questions on conflict prevention actors' methods and ethos: What does it mean for the United Nations to integrate forms of automated and predictive behavioral surveillance in intelligence collection and conflict analysis? What implications would "prevention as automated behavioral surveillance" have on human rights and the impartiality of peacekeepers? What normative vision and theory of harm do UN peacekeeping actors need to develop as they

67 Pauwels, "New Geopolitics of Converging Risks."

adapt to new technologies? Whose duty is it to foresee the unintended consequences of converging technologies on conflict-affected societies?

To ensure that the multilateral system is able to address 21st century crises and related conflicts, the United Nations and its member states will need to think creatively and plan for the risks and opportunities posed by converging technologies to their violence and conflict prevention work. Moving forward, they must exert robust normative leadership, partnering with the next generation of civil society and private sector actors to empower populations across the world.

The United Nations and its member states should

Bolster multistakeholder engagement by

- 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 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.

Develop a theory of harm by

- Formulating operational policy safeguards that ensure that the use of AI and converging technologies in violence and conflict prevention is in full compliance with international law. A theory of harm would involve developing adequate epistemic and normative methods to weigh the benefits of harnessing converging technologies, in particular automated and predictive behavioral monitoring, against the costs to civilian security and human rights.
- Relying on contractual, technical, and organizational mechanisms to ensure that sensitive dual-use technologies remain in the hands of strategic actors in the humanitarian and conflict prevention sectors and do not spread to organizations that fall outside the scope of due diligence.
- Increasing efforts to employ diverse expertise in the design, development, and testing of AI and converging technologies for violence and conflict prevention to ensure a broad range of perspectives and understanding of potentially sensitive use cases.
- Developing and operationalizing phase-gate processes that allow and require technological firms, policymakers, and actors in the humanitarian and conflict prevention sectors to consider the negative consequences and potential misuses of converging technologies during the full data life cycle, as well as during development, deployment, and postdeployment.
- Adhering to the purpose limitation principle in the collection, storing, processing, retention, and sharing of data in the violence and conflict prevention context. States and private sector actors will have to manage the tension between the opportunity to apply algorithmic analytics to mass population data and the privacy principles of data minimization and proportionality in intelligence collection.
- 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.

Develop human rights impact assessments by

- Conducting ongoing, in-depth human rights impact assessments for cases of potentially sensitive AI use and converging technologies in violence and conflict prevention. These assessments should focus not only on the data life cycle, of populations but also on design, development, and deployment. Cross-sector collaborations to anticipate unforeseen misuses and long-term impacts of AI and data capture technologies on vulnerable populations will be essential to ensure accountability.
- 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.

- Engaging with an appropriate range of external stakeholders, including nongovernmental organizations, researchers, and advocacy groups, with meaningful geographic diversity to ensure that their feedback informs UN, humanitarian, and private sector applications from the outset.
- Involving domain experts, including human rights domains, in the design process and operation of the AI systems used to make consequential decisions on the harnessing of population data sets.
- Maintaining robust feedback mechanisms on the performance of AI and converging technologies so they can be independently evaluated in terms of their predictive value, scientific validity, efficacy, and potential for bias, failure, and misuse.

Establish accountability mechanisms and remedy by

- Assessing the remediability of potential harms when deploying converging technologies and prioritize due diligence and transparency for those that are more challenging to rectify. In cases of frail or nonexistent data protection and redress mechanisms, disproportionate access to population data by state and private sector actors may create new power asymmetries. UN actors and partners should anticipate such asymmetries and accountability gaps by, inter alia, developing mechanisms for mitigating critical data incidents that would support a common theory of harm.
- Developing, implementing, and supporting effective grievance mechanisms for vulnerable populations whose rights may be harmed by converging technologies, including national civil and military justice systems and international or regional human rights mechanisms and courts.

ABOUT THE AUTHOR

Eleonore Pauwels

Eleonore Pauwels is a Senior Fellow for the Global Center on Cooperative Security. Her research focuses on security, governance, and ethical implications generated by the convergence of artificial intelligence (AI) with other dual-use technologies, including cybersecurity, genomics, and genome-editing. She regularly consults for the World Bank, the United Nations, governments, and private sector actors on AI and cybersecurity, the changing nature of conflict, foresight, and global security and counterterrorism.

ACKNOWLEDGMENTS

The Global Center gratefully acknowledges the support for this policy brief provided by the government of Norway. The views expressed are those of the author and do not necessarily reflect those of the Global Center or its advisory council, board, or sponsors.

ABOUT THE GLOBAL CENTER

The Global Center works to achieve lasting security by advancing inclusive, human rights–based policies, partnerships, and practices to address the root causes of violent extremism. We focus on four mutually reinforcing objectives:

- Supporting communities in addressing the drivers of conflict and violent extremism.
- Advancing human rights and the rule of law to prevent and respond to violent extremism.
- Combating illicit finance that enables criminal and violent extremist organizations.
- Promoting multilateral cooperation and rights-based standards in counterterrorism.

Our global team and network of experts, trainers, fellows, and policy professionals work to conduct research and deliver programming in these areas across sub-Saharan Africa, the Middle East and North Africa, and South, Central, and Southeast Asia.

GLOBAL CENTER ON COOPERATIVE SECURITY

globalcenter.org
@GlobalCtr

NORWEGIAN MINISTRY OF FOREIGN AFFAIRS