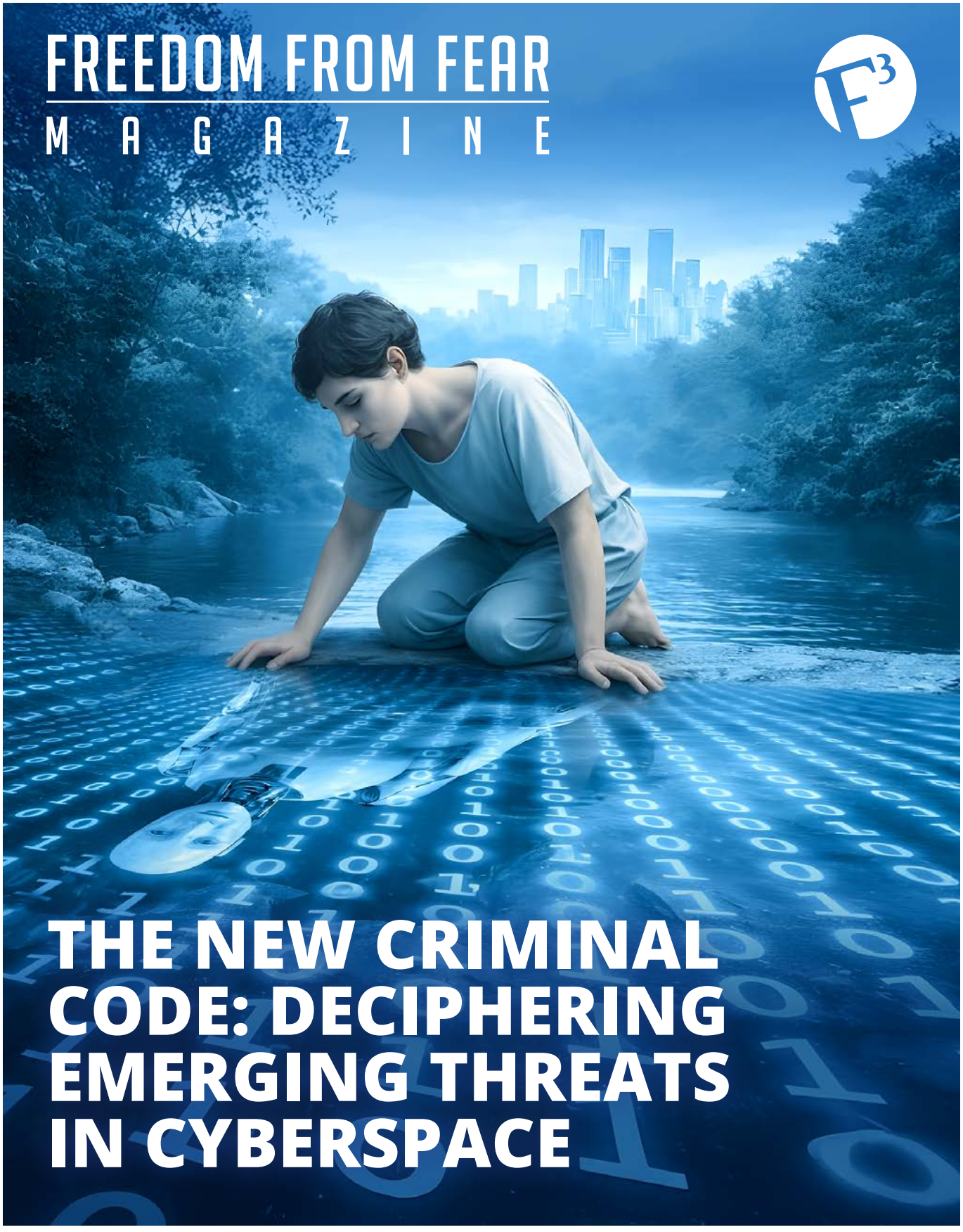
FREEDOM FROM FEAR

MAGAZINE



THE NEW CRIMINAL CODE: DECIPHERING EMERGING THREATS IN CYBERSPACE

unicri
United Nations
Interregional Crime and Justice
Research Institute

GHENT
UNIVERSITY

EDITORIAL BOARD

N.20

Disclaimer

FREEDOM FROM FEAR
MAGAZINE

THE NEW CRIMINAL CODE: DECIPHERING EMERGING THREATS IN CYBERSPACE

unicri
United Nations
Interregional Crime and Justice
Research Institute

GHENT
UNIVERSITY

CONTENTS

When code shapes crime: rethinking responsibility in the digital age

by Leif Villadsen
Acting Director of UNICRI

Foreword

Digital technologies are transforming the way societies function, communicate and protect themselves. At the same time, they are reshaping the nature of crime, creating new opportunities for harm while challenging existing legal, institutional and governance frameworks. Understanding emerging criminal threats in cyberspace has therefore become an essential priority for policymakers, practitioners and researchers alike.

Cyberspace is no longer simply a domain in which crime occurs; it is an environment that actively influences how criminal activities are organized, scaled and concealed. A wide range of digital technologies - from automated tools and data-driven systems to advanced online platforms - can enhance efficiency and connectivity, but they can also be exploited to lower barriers to entry, expand the reach of criminal networks and weaken traditional forms of oversight and accountability.

Criminal actors have demonstrated a notable capacity to adapt to these changes. Organized crime groups increasingly operate across digital and physical spaces, exploiting online infrastructures to facilitate cyber-enabled crime, fraud, illicit financial flows and other forms of transnational criminal activity. Terrorist actors similarly use digital en- vironments to disseminate propaganda, support recruiment and operate in decentralized ways that are more difficult to detect and disrupt. These dynamics highlight how emerging technologies are reshaping crime in cyberspace, often faster than regulatory and institutional responses can adapt.

At the heart of these challenges lies a fundamental tension between technological innovation and human responsibility. Digital systems, including those based on artificial intelligence, can process vast amounts of information and support decision-making, but they do not possess judgement, context or ethical awareness. Machine "reasoning" is based on mathematical models, statistical correlations and optimization towards predefined objectives. The risks associated with these developments extend beyond technical harm. Autonomous and semi-autonomous systems, whether deployed for security, surveillance or operational purposes, cannot make ethical judgements. A system may identify a threat or a target, but it cannot assess proportionality, intent or human dignity. When such technologies are misused or insufficiently governed, they risk eroding accountability, undermining trust and blurring the line between human decision-making and algorithmic action.

This is where human-trust solutions play a vital role. Keeping human beings meaningfully involved in decision-making processes, ensuring transparency and contestability of technological systems, and clearly defining responsibilities are essential elements of effective responses to cybercrime and online harms. Preserving digital trust is not only a technical objective, but a strategic and societal imperative.

UNICRI's work on cybercrime and online harms contributes to this effort by supporting States and stakeholders in strengthening prevention, enhancing criminal justice responses and promoting international cooperation in the digital space. Through research, capacity-building and policy-oriented initiatives, UNICRI seeks to address the evolving risks posed by cyber-enabled and cyber-dependent crime, while reinforcing respect for human rights and the rule of law.

This issue of F3 marks the first in a series dedicated to these themes. It explores the dynamics shaping the new criminal code of the digital era, shedding light on emerging criminal threats in cyberspace and on how technological change is reshaping criminal behaviour.

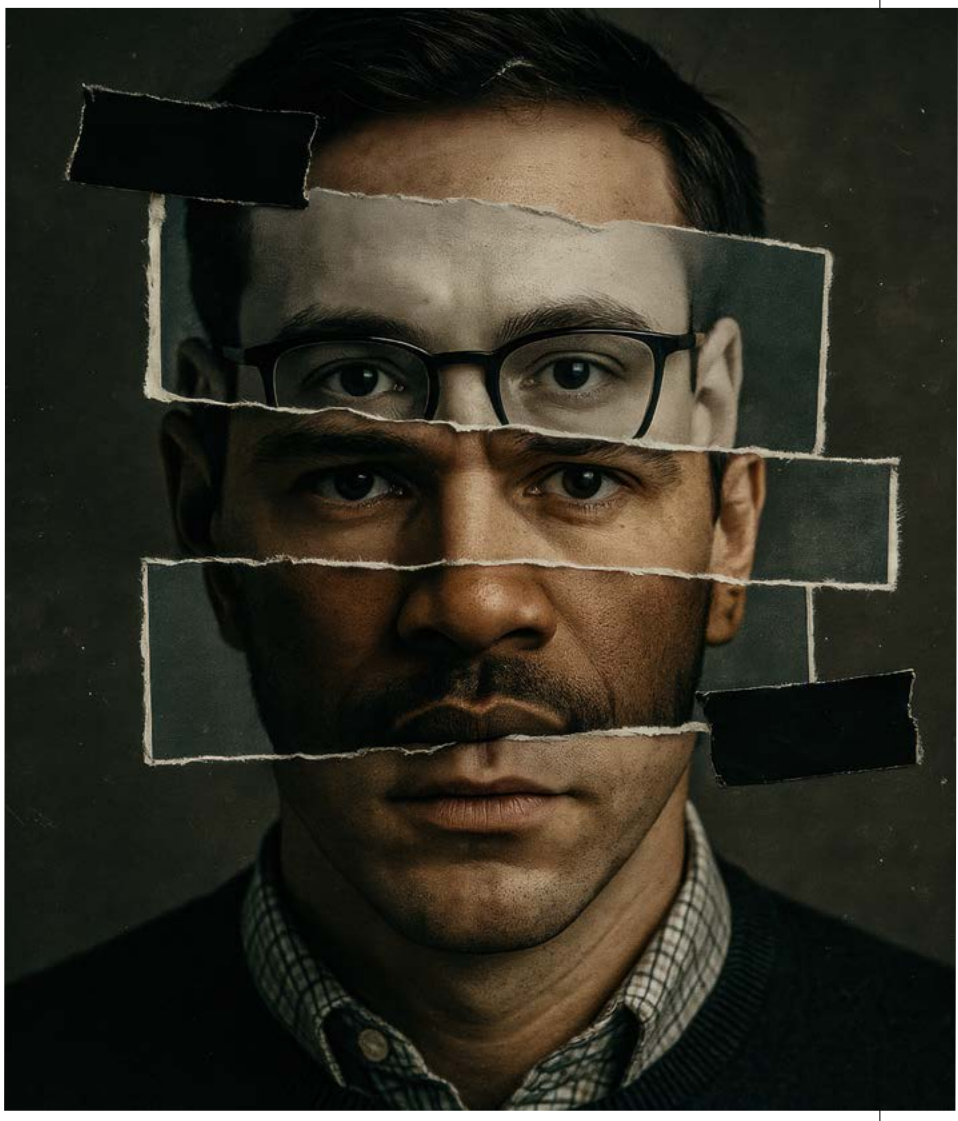
The issue gives voice to a diverse range of stakeholders, including young people, to reflect the many facets of the challenges posed by rapid tech- nological advancement for security and human rights, as well as the responses that are possible. It underscores that prevention and response must be inclusive and collaborative, involving all relevant actors - from governments and academia to the private sector and civil society.

By fostering informed dialogue and shared understanding, this issue aims to contribute to the collective effort to preserve digital trust and to promote an open, safe and secure digital era for all.

AI crime: what we know, what we don't know, and what we need to know

by Marie-Helen (Maria) Maras

Criminals have leveraged information and communications technology (ICT) to commit crime, with their targets, tactics, and methods of operation evolving to adapt to criminal justice measures. It is no surprise, therefore, that criminals have used emerging technologies such as artificial intelligence (AI) to improve the efficiency of their operations, expand their reach and impact, obscure their illicit activity, break barriers to entry into illicit markets, and commit various forms of crime. AI is (and/or can be) used by criminals to commit *crimes against the machine* (i.e., offenses against the confidentiality, integrity and availability of data, systems, and networks; cyber-dependent crime); *crimes using the machine* (i.e., computer-related offenses; cyber-enabled crime); *crimes in the machine* (i.e., content-related offenses; cyber-enabled crime);[1] and *crimes by the machine* (i.e. crimes committed by the AI).



Crimes against the machine

AI has served as a tool to commit cyber-dependent crime, such as hacking and malware creation and distribution. AI is used to develop and enhance malware, enabling it to operate quickly in identifying system vulnerabilities and ways to exploit these vulnerabilities.[2] AI is also used to create advanced forms of malware that adapt and evade security software detection. Criminals have committed cyber-attacks against AI by corrupting or poisoning data used to train large language models (LLMs) and AI *(data poisoning)* and manipulating AI data inputs *(malicious injections)*. Cracked versions of legitimate AI models known as *dark AI*, have been used for criminal

1 The categories referenced (crimes against the machine, crimes using the machine, and crimes in the machine) are drawn from David Wall's (2007; 2024) cybercrime typology.

2 Sancho, D. and Ciancaglini, V. (2023). Hype vs. Reality: AI in the Cybercriminal Underground. Trend Micro, August 15, 2023.

purposes as well.[3] While the capabilities of DarkAI applications were reported as limited in 2023,[4] in 2025, variations of DarkAI surfaced that were purportedly designed to overcome those limitations. Specifically, new evolutions of DarkAI (e.g., Xanthorox), which are designed for cybercrime purposes, are purportedly custom-built, self-sustaining language models that are hosted on private servers controlled by the developer; this design, coupled with its offline functionality, makes it harder to trace, detect and ultimately take down.

Crimes using the machine

AI has served as a tool to commit cyber-enabled crime, such as fraud and technology-facilitated violence. *Jailbreaking* (i.e., disabling default software limitations) and *prompt engineering* (i.e., instructions designed to achieve desired results) have been used to make commercial AI applications (e.g., ChatGPT) act in unanticipated and harmful ways. In cyber-enabled fraud *cases,* these tactics have been used to create fake websites (pharming or phishing websites) designed to steal targets' personal information, or to create scripts to facilitate various forms of cyber-enabled fraud, such as hybrid investment fraud, "whereby criminals gain the trust of victims by

3 Poireault, K. (2023). The Dark Side of Generative AI: Five Malicious LLMs Found on the Dark Web. Infosecurity Europe, August 10, 2023.

4 Sophos X-Ops. (2023). Cybercriminals can't agree on GPTs. Sophos News, November 28, 2023.

forming connections and relationships, and then exploit[ing] this trust ...to [get] victims to... invest in securities or commodities."[5] Moreover, there have been several instances where *deepfakes*, hyper-realistic media that depict a person saying or doing something they did not say or do, have supposedly been used to facilitate various forms of cybercrime.

Deepfakes have been used to create real or synthetic identities to facilitate other forms of fraud, even real estate

AI-manipulated audio (or voice cloning) and/or video has been reported in certain cyber-enabled fraud cases. In cases where AI-manipulated audio was suspected, victims of various forms of cyber-enabled fraud claimed to have heard the voice of their loved ones in duress, screaming and/or asking for help. For example, in so-called grandparent frauds, a target (a grandparent) purportedly receives a distressing phone call from or relating to their grandchild requesting money to help them in a situation they find themselves (e.g., accident, jail, kidnapping, etc.). Furthermore, deepfakes have been used to create real or synthetic identities to facilitate other forms of fraud, even real estate fraud by posing as homeowners to sell property that does not actually belong to the imposters.[6] Finally, criminals have used AI, specifically generative AI, to manipulate media (image, audio and video) to facilitate numerous forms of technology-facilitated violence, including cyber harassment, sextortion, and image-based sexual abuse (or non-consensual dissemination of intimate media such as images).[7]

Crimes in the machine

AI-developed content, such as deepfakes, could be criminal if what is depicted is considered illegal. For instance, criminals have used AI to generate child sexual exploitation material (CSEM) and child sexual abuse material (CSAM). In 2025, in Operation Cumberland, which involved 19 countries, numerous individuals linked to an online platform dedicated to the creation and distribution of AI-generated CSAM were identified and arrested.[8] In jurisdictions where image-based sexual abuse – a form of technology-facilitated violence that predominantly targets *women* – is considered illegal, non-consensual nude and sexually explicit deepfakes or other forms of manipulated media would be considered illegal.

Crimes by the machine

This final category of AI crime is an anticipated evolution of crimes committed by autonomous AI systems. The commission of these forms of crime poses particular challenges to the criminal justice system, including the determination of accountability and liability for these crimes and effective remedies.

5 Maras, M.-H. and Ives, E. (2024). Deconstructing a Form of Hybrid Investment Fraud: Examining 'Pig Butchering' in the United States. *Journal of Economic Criminology*, 5 (Special Issue on Relationship Fraud: Romance, Friendship and Family Frauds), 100066.

6 Walker, H. and Hall, D. (2024). 'I don't think any property is safe': South Florida man says he almost became a victim to AI real estate fraud. *WSVN 7News*, September 19, 2024.

7 O'Brien, W. and Maras, M.-H. (2024). Technology-Facilitated Coercive Control: Response, Redress, Risk, and Reform. *International Review of Law, Computers and Technology* (Special Issue on Digital and Online Violence), 38(2), 174–194; Maras, M.-H. and Logie, K. (2024). Countering the Complex, Multifaceted Nature of Deepfakes: An Augean Task? *Crime Science* (Special Collection on Measuring, Detecting, and Preventing Cyber Social Threats), 13(31), 1-17.

How pervasive is this?

The short answer is: we don't really know. Cyber organized criminal groups, and other offenders have used AI to commit various cybercrimes. While these incidents shed some light on the types of cybercrimes committed using AI and how the AI was used to commit cybercrime (including, where information is available, the targets and perpetrators of these crimes as well as the types of tools used in the commission of these crimes), they do not provide information about the true nature and extent of AI crime worldwide.

Standardized data collection and reporting mechanisms on AI crime do not exist worldwide. The categories of AI crime, types of data collected about these crimes, and criteria used to measure these crimes also do not exist. In short, apart from criminal cases and incidents reported in government press releases and news reports, information about AI crime is not available. Comprehensive and harmonized AI crime data collection and recording practices are needed to identify the nature and extent of AI crime. This data is also needed to inform the public, policies, and measures to control, reduce, mitigate, and prevent AI crime.

Comprehensive and harmonized AI crime data collection and recording practices are needed to identify the nature and extent of AI crime

To capture this data, an initial first step involves the mutual understanding of AI crime within and across countries by standardizing AI crime definitions and categories. The lack of standardization and harmonization inhibits the accurate capture of information about AI crime (including information about offenders and targets of this crime) and leads to a dearth in available (and accurate) AI crime data. This data, if collected in a uniform manner and made available for analysis, could reveal AI crime prevalence, patterns and trends (such as the modus operandi of illicit actors, emerging crimes, specific types of crimes occurring more frequently), which can be used to inform the development of targeted interventions and create evidence-based policies and measures to counter AI crime.

ABOUT THE AUTHOR

Dr. Marie-Helen (Maria) Maras is a tenured Professor with a joint appointment in the Department of Sociology and the Department of Mathematics and Computer Science and the Director of the Center for Cybercrime Studies at John Jay College of Criminal Justice, City University of New York. Her education is multidisciplinary, covering law, criminology, criminal justice, psychology, and computer and information science. Her academic background and research focus on the evolution of transnational crime and the legal, ethical, social, and political impact of digital technology, especially emerging technologies. She has published numerous peer-reviewed academic journal articles and books, including Real Criminology (Oxford University Press), Cybercriminology (Oxford University Press), and Computer Forensics: Cybercriminals, Laws, and Evidence (Jones and Bartlett), among other publications. Dr. Maras has engaged in counter cybercrime capacity building activities as a consultant for the United Nations Office on Drugs and Crime (UNODC), the Organization for Security and Co-operation in Europe (OSCE), the International Development Law Organization (IDLO), and the Inter-American Development Bank (IDB).

ISSUE 5: Soon Online!

unicri
United Nations Interregional Crime and Justice Research Institute

1540 Compass

ISSUE FIVE: INTERVIEWS WITH

PERMANENT REPRESENTATIVE OF PANAMA TO THE UN

Eloy Alfaro de Alba

DIRECTOR, SECRETARIAT OF THE ASIA-PACIFIC GROUP ON MONEY LAUNDERING

Mitali Tyagi

FORMER COORDINATOR OF THE 1540 GROUP OF EXPERTS

Jonathan Brewer

Breaking the machine: jailbreaking AI and safeguarding the digital frontlines of vulnerable communities

by Fabien Leimgruber and Alexandru Lazar

Artificial intelligence is no longer confined to tech firms. In the nonprofit world, AI is increasingly developed and woven into the daily work of organizations serving vulnerable communities - allocating humanitarian aid, managing beneficiary data, monitoring and tracking project delivery, or coordinating emergency response. Many of these systems are hosted in the cloud, built by external third-party providers, and designed to process highly sensitive personal and behavioral data. As AI's capabilities expand, so does the ambition of those looking to exploit it. Jailbreaking has emerged as an accessible technique for bypassing safeguards, extracting hidden data, and manipulating outputs.[1] Combined with zero-day vulnerabilities, these attacks can undermine the very integrity of the digital infrastructure on which critical social services depend. A zero-day vulnerability in an AI system or tool is an undiscovered flaw in the system's code, model, or supporting infrastructure that threat actors can exploit before the organization or vendor is aware of it or able to patch it.[2] A jailbreak can uncover hidden weaknesses in the system's logic, giving malicious actors a way to exploit them. For the nonprofit sector, such security breaches are not abstract risks - they can translate directly into physical danger.

How jailbreaking differs from other cyber threats

Jailbreaking attacks the logic of the AI model itself. Instead of exploiting code or network weaknesses and breaking into a server, adversaries work to bend the system's understanding of language, context, and rules. This can mean changing prompts in ways that slip security filters, embedding hidden instructions inside seemingly harmless text, or using languages and dialects that evade existing safeguards. Once a method is discovered, it can be shared and replicated with remarkable speed. And unlike a fixed piece of code, AI models' vulnerabilities can shift depending on the data they process or the con-

1 Kranz, T. & Jonker, A. (2025), AI jailbreak: rooting out an evolving threat, IBM. https://www.ibm.com/think/insights/ai-jailbreak

2 Shastri, V. (2025), What is a zero-day exploit?, Crowdstrike. https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/zero-day-exploit/

text in which they operate. The risks go far beyond producing offensive or unsafe outputs. Jailbroken systems can be coerced into revealing proprietary prompts, leaking training data, or exposing bias patterns that adversaries can then exploit.

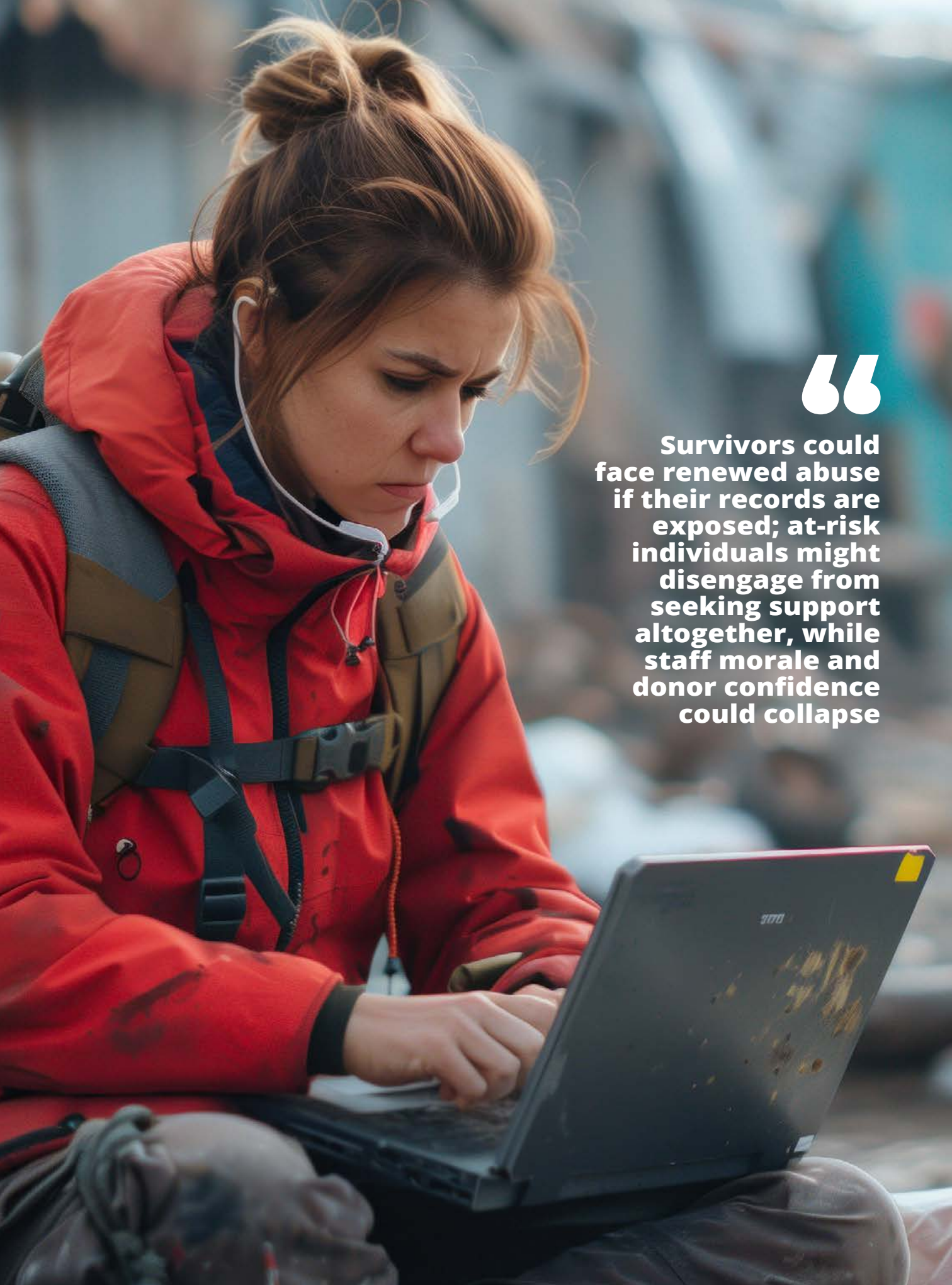
In 2023, a wave of public "persona" jailbreak prompts, most notably the "DAN" (Do Anything Now) prompt, spread rapidly across online communities. These prompts instructed conversational AI models like ChatGPT to assume an alternate persona supposedly freed from their normal content restrictions. By framing requests within this role-play, users were able to elicit disallowed or sensitive information that the models would ordinarily refuse to produce.
This form of jailbreak, driven by social prompt engineering rather than technical exploitation, demonstrated how easily safety filters could be bypassed through clever manipulation of language.

It led to numerous documented incidents in which AI systems generated harmful, misleading, or prohibited content despite embedded safeguards.[3]

Why nonprofits are uniquely exposed

A large corporation hit by an AI breach may suffer reputational damage and financial loss, but it often has dedicated security teams and the resources to recover quickly and resume normal business. Nonprofits do not enjoy such resilience and business continuity. They frequently rely on donated or discounted tools that they cannot fully configure or monitor. Mission urgency often drives rapid adoption of new technologies, leaving little room for deep security vetting. Threat actors, and in particular

3 Taylor, J. (2023), ChatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards, The Guardian. https://www. theguardian.com/technology/2023/mar/08/ chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards



Survivors could face renewed abuse if their records are exposed; at-risk individuals might disengage from seeking support altogether, while staff morale and donor confidence could collapse

cybercriminals, see nonprofits as rich targets for three intertwined reasons:

1. Because of what they do: nonprofits often fill crucial gaps in social services, and as visible agents of change, they are frequent targets for actors who wish to discredit or neutralize them.
2. Because of what they know: these organizations collect and manage sensitive data about vulnerable populations, local networks, and operational plans, making them attractive to attackers seeking to steal, extort, or weaponize that information.
3. Because of what they have: nonprofits do raise funds, yet these funds are restricted for project delivery providing lifesaving support to millions of vulnerable individuals.

Only a very small part of a nonprofit's budget is invested in cybersecurity personnel and tools.[4] Therefore, this combination of high-value data, limited control, and stretched resources creates a dangerous gap. Since 2018, there have been over 590,000 threats impacting the nonprofit sector.[5]

As an example, consider a nonprofit operating a digital counseling platform for survivors of gender-based violence. To extend its reach, the organization deploys a custom AI system that screens initial requests, offers guided self-help resources, and directs urgent cases to human counselors. The model is trained on sensitive intake data, including narratives of abuse, geolocation details, and mental health assessments. A jailbreak attack against this system could unfold in several ways:

1. By tricking the AI into revealing snippets of its training data, an adversary could extract private testimonies, placing survivors at risk of being re-identified or located.
2. Threat actors might coerce the model into delivering

4 CyberPeace Institute (2023), CyberPeace Analytical Report:NGOs serving Humanity at risk: Cyber Threats affecting "International Geneva". https://geneva.cyberpeace.ngo/

5 CyberPeace Institute (2025), CyberPeace Tracer. https://cyberpeacetracer.ngo/

harmful instructions - encouraging self-harm instead of coping strategies, for example - undermining trust and endangering lives.

3. A successful attack could disable triage logic, flooding human counselors with false 'urgent' cases and delaying help for those in genuine crisis.

For the nonprofit sector, as in this example, the consequences extend far beyond reputational damage. Survivors could face renewed abuse if their records are exposed; at-risk individuals might disengage from seeking support altogether, while staff morale and donor confidence could collapse.

"

Since 2018, there have been over 590,000 threats impacting the nonprofit sector

Threat detection, policy and responsibility in the AI era

Defending against jailbreaking requires tools and methods different from traditional network monitoring. It means stress-testing models through continuous prompt injection simulations (deliberate attempts to trick the AI into ignoring its safeguards), building red-teaming exercises specific to AI (structured tests to mimic attacker strategies), and watching for patterns in model behavior that may indicate tampering.

For nonprofits, this is rarely achievable alone. As such, there is a pressing need for:

- Secure communication channels for sharing jailbreak techniques and alerts on AI zero-day vulnerabilities, as well as mitigation strategies across trusted partners, both from the private tech sector and nonprofit organizations.

- Safe spaces where nonprofit organizations can test their models against advanced attacks without risking live data.

- Investment in staff AI literacy to bridge mission delivery with AI risk awareness, even if they are not full-time security professionals.

Without these measures, the nonprofit sector risks falling further behind in a rapidly evolving threat landscape. In this context, the CyberPeace Builders program from the CyberPeace Institute is a vital initiative connecting nonprofits with pro bono cybersecurity and AI experts who provide practical support, such as cyber hygiene training, AI security awareness, and technical assistance.[6] Through this initiative, organizations can also access tailored resources on AI skills training for staff, AI responsible use, and policy templates - helping them strengthen digital resilience without diverting scarce funds from their missions.[7] Moreover, technical solutions are only part of the answer. Procurement policies must explicitly require AI security reviews, data minimization practices, and clear incident response plans that also include AI threat scenarios.

> **In cybersecurity, reacting is costly; anticipating is essential**

Since AI systems are built and deployed across borders, harmonizing standards internationally is essential - so that a frontline nonprofit in Nairobi benefits from the same safeguards as one in New York. Technology providers must also bear responsibility. The nonprofit sector should not be left with "lite" or "free" versions of AI tools that strip away critical protections. Providing secure, fully-featured systems should be considered a core ethical obligation.

Staying ahead

In cybersecurity, reacting is costly; anticipating is essential. For nonprofits working with sensitive data, staying ahead of AI jailbreak techniques is part of their duty of care. That means ongoing investment in staff training, governance, and partnerships that cut across sectors and borders. Jailbreaking and exploiting potential zero-day vulnerabilities in nonprofit AI systems is a reminder that, as these tools become smarter, defenses must become sharper.

In spaces where digital compromise can cause real-world harm, particularly to those least able to protect themselves, there is no room for complacency.

Security here is not just about protecting systems - it is about protecting lives.

ABOUT THE AUTHORS

Fabien Leimgruber is the Head of Cyber Resilience at the CyberPeace Institute. He joined the Institute in 2020, where he works to strengthen the cyber resilience of public-interest organizations and oversees the expansion of the CyberPeace Builders programme. Fabien has experience across the public, private, and nonprofit sectors. He previously led the Cyber Threat Intelligence Unit at Kudelski Security and later served as Information Security Awareness Adviser at the ICRC. He holds a Master of Law in Criminality and Information Security from the University of Lausanne.

Alexandru Lazar is a Nonprofit Cyber Resilience Specialist at the CyberPeace Institute, which he joined in 2021. At the Institute, he supports nonprofits in strengthening their cybersecurity resilience and helps coordinate the CyberPeace Builders, a network of corporate volunteers assisting public-interest organizations. He holds an Erasmus Mundus International Master's in Security, Intelligence and Strategic Studies from the University of Glasgow, as well as an MA in Politics and International Relations from the University of Aberdeen.

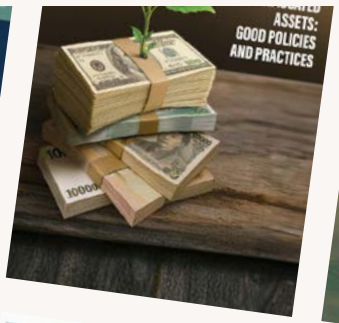
6 CyberPeace Institute (2025), CyberPeace Builders https://cpb.ngo/

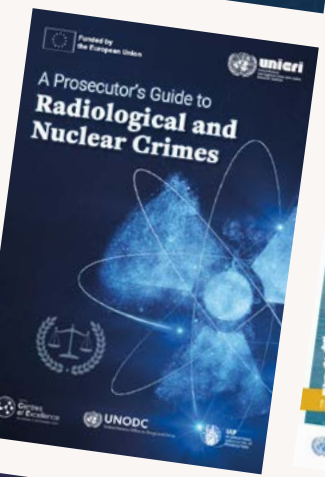
7 CyberPeace Institute (2025), AI Skills for Nonprofits: A Collective Effort. https://cyberpeaceinstitute.org/news/ai-skills-for-nonprofits-a-collective-effort/

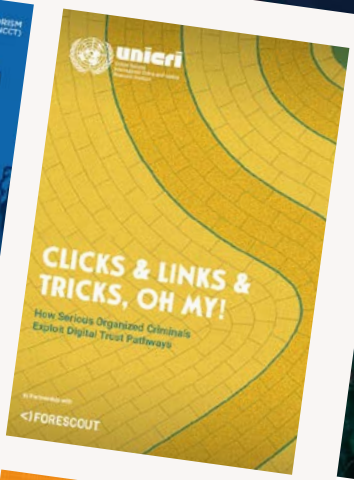
unicri
United Nations
Interregional Crime and Justice
Research Institute

Publications

Download UNICRI publications







Generative AI: a new threat for online child sexual exploitation and abuse

by Yalda Aoukar, Lisa Maddox and Lana Apple

"[He] took that cherished memory and turned it into a new memory — one that elicits nausea, fear, and overwhelming discomfort and distrust within me."[1]

This 40-year-old woman was speaking at the sentencing of her former classmate (more recently, a North Carolina child psychiatrist), who had used artificial intelligence (AI) to turn her innocuous childhood photos into sexual abuse material. Just a few years ago, this crime would have been impossible; today, it is alarmingly on the rise.

In this case, the offender resurfaced old images, including one from her first day of school: she and a group of 15-year-old students stood smiling and waiting for the bus. Using AI, the offender transformed that childhood moment into explicit images.

The offender was convicted under U.S. law of producing, transporting, and possessing child sexual abuse material (CSAM), which has been expanded to include AI-generated images.

While this was the first case of its kind in North Carolina, the world is quickly having to grapple with this new threat. This year, Europol announced a sweeping operation against a global network producing and distributing AI-generated CSAM.

1 FBI.gov, "'Horribly twisted' Charlotte pornography case shows the 'unsettling' reach of AI-generated imagery," April 2024.

Authorities identified 273 suspects, arrested 25 indivi duals, and carried out 33 house searches. The true scope is likely much larger: the group's platform allowed anyone with a password to log in and watch AI-generated material of children being abused, connecting offenders across the globe in real time.[2]

Online sexual exploitation and abuse have been endangering children across the world since the invention of the Internet, with one out of eight children around the world falling victim to online sexual exploitation.[3] The production, possession, and distribution of CSAM is a significant component of online child sexual exploitation and abuse, resulting in long-term harm to victims. CSAM includes any representation of a child engaged in real or simulated explicit sexual activities and threatens children across the globe.

The rise of AI is altering the existing landscape of child harm. One of the most pressing dangers facing the global child protection ecosystem is AI's effects on CSAM. Of the approximately 29.2 million incidents of online child exploitation and abuse reported to the National Center for Missing & Exploited Children (NCMEC)'s Cyber Tipline in 2024, 67,000 included verified AI-generated CSAM.[4]

The group's platform allowed anyone with a password to log in and watch AI-generated material of children being abused, connecting offenders across the globe in real time

While this number is still comparatively low, experts anticipate a stark increase as the capabilities to create AI-generated CSAM advance.[5] In fact, between 2023 and 2024, NCMEC saw a 1,325% increase in reports involving generative AI.[6] Understanding the landscape of AI-generated CSAM is a necessary precursor to safeguarding children.

To enable a shared understanding of the current landscape of AI-generated CSAM, the Bracket Foundation, the United Nations Interregional Crime and Justice Institute (UNICRI), and Value for Good collaborated on the report "Generative AI: A New Threat for Online Child Sexual Exploitation and Abuse"[7], which pulled together perspectives and data from law enforcement, technology companies, civil society, and caregivers, and provided a comprehensive overview of the escalating danger and suggested potential mitigation strategies for improving the safeguarding of children. The insights in the report were derived through qualitative and quantitative research methodologies:

- **Aggregation of existing data and publications:** review and collection of existing published data and literature.
- **News reporting:** thorough review of news reports on the topic of AI-generated CSAM.
- **Expert interviews:** interviews with 17 experts from the public and private sectors, law enforcement, and civil society.

2 Europol, "25 arrested in global hit against AI-generated child sexual abuse material," February 2025

3 Childlight Global Child Safety Institute, "Into the Light Index"; Andy Gregory, "Prioritise Children's Online Safety at Election to Tackle 'Hidden Pandemic' of Sexual Abuse, Experts Urge," The Independent, June 2024.

4 John Shehan, "Addressing Real Harm Done by Deepfakes", NCMEC, March 2024.

5 NCMEC, "2024 CyberTipline Report," 2024.

6 NCMEC, "2024 CyberTipline Report." 2024.

7 Bracket Foundation, UNICRI, Value for Good, "Generative AI: A New Threat for Online Child Sexual Exploitation and Abuse," September 2024.

- **Internal investigations and analysis:** analysis and review of legislation and policy, and open access data sources, as well as original data collection and analysis of surveys of law enforcement and caregivers. Two anonymous surveys were conducted by Value for Good in Spring 2024 for inclusion in this report (law enforcement: 107 officers in 28 countries; caregivers: 103 respondents from 15 countries). While the survey results make no claim of global representation, they do give an indication of the major challenges facing these two stakeholder groups in addressing AI-generated CSAM.

In this context, the key to understanding AI-generated CSAM is to realize that it is more than just computer-generated imagery. Current forms of AI-generated CSAM include:

- **Text content:** Generative AI chatbots have been shown to engage in sexually explicit chats, acting as children might, and generative AI has also generated guides, tutorials, and suggestions on how to sexually abuse children.[8]
- **Still and moving imagery:** Generative AI models are increasingly able to generate photorealistic CSAM and alter existing imagery to make it explicit. As this technology improves, perpetrators can create higher-quality moving imagery and videos.

Crucially, generative AI has also expanded how children - and adults, like in the North Carolina case - can become victims. The wide range of victims includes:

8 Testimony of John Shehan, "Addressing Real Harm Done by Deepfakes", March 2024.

- Children whose innocuous images have been used to train AI models.
- Children whose innocuous images are transformed into CSAM with AI.
- Existing victims of CSAM, who have been revictimised through the modification or obscuration of existing CSAM.
- Adults whose images have been de-aged to create AI-generated CSAM.

In addition to adult perpetrators of child sexual abuse, young people themselves are increasingly becoming creators of AI-generated CSAM through the use of 'nudify' apps. 'Nudify' or 'undressing' apps use AI to 'undress' images, predominantly of women and girls. By filling in a 'best-guess' of what the woman in the picture would look like without her clothing on, 'nudify' apps create non-consensual nude images. As of 2024, U.K. law enforcement estimated that at least one child in every school in the United Kingdom has one of these apps. While using these apps on images of children may not always meet the definition of AI-generated CSAM, they are still a serious and increasingly widespread danger.[9]

To confront this myriad of rapidly evolving threats, a coordinated, cross-sector response is no longer optional; it is urgent. As Simon Bailey, Director of Strategic Engagement at the Child Rescue Coalition, remarked: "We are at the point that highly motivated offenders can take an image and do whatever they want with it". Combating this new reality requires every relevant actor to step up, work together, and close the gaps that offenders exploit.

Generative AI chatbots have been shown to engage in sexually explicit chats, acting as children might, and generative AI has also generated guides, tutorials, and suggestions on how to sexually abuse children

We are calling on stakeholders to understand the threat and act accordingly:

- **Private Sector:** AI developers must implement safety measures to prevent models from generating explicit content, particularly involving children. Technology platforms should prioritize children's safety by blocking and moderating AI-generated CSAM, cutting distribution channels.
- **Governments:** Policymakers must update laws to address AI-generated CSAM, requiring systemic reforms and increased investments in technology. Collaborations with technology providers are essential to ensure robust child safeguards.
- **Law Enforcement:** Agencies need to stay updated on AI-generated CSAM trends through international exchanges and adopt new tools for identifying such content, ensuring effective responses.
- **Caregivers:** Parents and guardians must stay informed about online threats to children, openly discuss Internet dangers, and utilise available resources. Limiting children's online presence should also be considered.

9 Bracket Foundation, UNICRI, Value for Good, "Generative AI: A New Threat for Online Child Sexual Exploitation and Abuse," September 2024.

ABOUT THE AUTHORS

Yalda Aoukar is Co-Founder and Managing Partner of Bracket Capital and President of Bracket Foundation, which leverages technology for social good with a focus on online child safety and addressing global challenges. She sits on the boards of the UN's AI for Safer Children Initiative and the World Innovation Summit for Education Accelerator and was a 2024 Young Global Leader of the World Economic Forum.

Lisa Maddox is a Principal at Value for Good, a Berlin-based social impact consultancy, focusing on Tech for Good and co-leading the Global Development practice. She previously served as Chief of Staff at Fuzu in Nairobi and worked as a project leader at Accenture in New York.

Lana Apple is a Senior Consultant at Value for Good specializing in education, child rights, and international development. She holds a master's degree from the University of Oxford, has taught in the United States and Germany, and has published research on global education issues.

Generative artificial intelligence for human rights: disregarded prosocial uses of deepfake technology and their connection to human rights

by Can Yavuz and Gert Vermeulen

Introduction

Nowadays, it is difficult to escape conversations about generative artificial intelligence (GenAI). This widespread attention stems from its rapid advancement. A glance back at the dawn of the century demonstrates the scale of this progress. In 2002, the science fiction movie *S1m0ne* portrayed the story of a producer who created a digital actress to substitute the lead actress who had walked away from his film. That same year, in *Ashcroft v. Free Speech Coalition*[1], the Supreme Court of the United States struck down a law that criminalised virtual child sexual abuse material on the grounds that it lacked photorealism. In the dissenting opinion, Justice Rehnquist warned that rapidly advancing technology would soon make computer-generated images indistinguishable from authentic ones. That day has arrived. Today, GenAI systems can create synthetic but hyperrealistic images, videos, and audio, referred to as deepfakes.

How society uses and regulates a disruptive technology is considerably influenced by its public perception. In the case of deepfake technology, media coverage and academic discussions have substantially concentrated on its misuse,[2] often presented through eye-catching headlines. "AI could set us back 100 years when it comes to how we consume news[3]", "Will deep-fake technology destroy democracy?[4]", and "AI-assisted fake porn is here and we are all f**ked[5]" were some of the many examples. This narrative has fostered the negative connotation of deepfake[6] and arguably reduced it to a threat against human rights.

There is no denying that deepfake technology has been used for malicious purposes. One of the first widespread (mis)uses of deepfakes was image-based sexual abuse, which was followed by its weaponization for disinformation, online child sexual abuse, fraud, and fabricated evidence. Nevertheless, focusing solely on the dark side of deepfakes presents an incomplete and potentially misleading picture. Deepfake technology has many (overlooked) prosocial uses that can help the realization of human rights. Given that the narrative surrounding a disruptive technology may influence its use and its regulation, it is necessary to look at the other side of the coin to gain a more nuanced understanding. Thus, let us explore the often disregarded prosocial uses of deepfake technology and their connection to human rights.

Prosocial uses of deepfake technology and their connection to human rights

The following explores the disregarded prosocial uses of deepfake technology and how they can contribute to the realisation of the freedom of expression, the right to education, the right to health, and the right to justice and security, respectively.

Regarding freedom of expression, deepfake technology unlocks opportunities for political speech. While behind bars, Pakistani politician Imran Khan used deepfake technology to deliver political speeches and

1 The Supreme Court of the United States, *Ashcroft v. Free Speech Coalition*, 535 U.S. 234, April 16, 2002.

2 Can Yavuz, "A Multidisciplinary Look at History and Future of Deepfake with Gartner Hype Cycle", *IEEE Security & Privacy*, Vol. 2, Issue 2024.03, (May/June 2024). Alexander Godulla and others, "Dealing with Deepfakes - An Interdisciplinary Examination of the State of Research and Implications for Communication Studies", *Studies in Communication and Media*, Vol. 10 Issue 1, (January 2021).

3 Jackie Snow, "AI Could Set Us Back 100 Years When It comes to How We Consume News", MIT Technology Review, 7 November 2017.

4 Jennifer Finney Boylan, "Will Deep-Fake Technology Destroy Democracy?", The New York Times, 17 October 2018.

5 Samantha Cole, "AI-Assisted Fake Porn is Here and We Are All F**ked, Vice, 11 December 2017.

6 Mariëtte van Huijstee and others, Tackling Deepfakes in European policy, (Brussels, European Parliamentary Research Service, 2021).

run a successful campaign.[7] Deepfake technology can also be instrumental in activism, exemplified by the *Unfinished Votes* project. It digitally brought Joaquin Oliver — a 17-year-old victim of one of the deadliest school shootings in the United States of America — back to deliver one last message. In his deepfake, Oliver asks people to replace his vote in the next election for strict gun control laws.[8] Turning to satire, which also enjoys protection under freedom of expression, a noteworthy example is the deepfake web series Sassy Justice. It created deepfakes of fictionalised public figures and placed them into the world of a local journalist. The deepfake-powered series covers complex topics like disinformation, media independence, and nepotism in an engaging and light-hearted fashion.[9] Deepfake technology has the potential to enhance artistic expression as well. While this tool lowers the entry barrier for creative expression and empowers those lacking artistic skills, a growing number of artists use GenAI to create art.

“

It digitally brought Joaquin Oliver — a 17-year-old victim of one of the deadliest school shootings in the United States of America — back to deliver one last message. In his deepfake, Oliver asks people to replace his vote in the next election for strict gun control laws

The right to education is another human right that can benefit from deepfake technology. This technology enables the smooth translation of auditory educational material across languages. Additionally, it can significantly lower the production costs of audiovisual educational content, making it more accessible and personalized for diverse learning needs. Deepfake technology can also provide a more interactive learning experience, particularly for visual learners. The Dalí Museum, which created deepfakes of Salvador Dalí, is an example of these capabilities. The deepfake Dalí tells the story of his artwork and takes a selfie with museum visitors, creating lasting impacts on visitors.[10]

Deepfake technology presents innovative opportunities for the right to health. This is particularly true for health data sharing and medical research. The synthetic nature of deepfake can facilitate data masking, thereby contributing to privacy-preserving health data sharing. In some instances, the realism of deepfake health data can serve as a viable substitute for authentic data and support data augmentation in medical research.[11] Additionally, deepfakes can improve the quality of life for individuals with rare diseases. As an assistive self-visualisation tool, deepfake technology can be a remedy for aphantasia patients (people who cannot voluntarily generate visual imagery). Moreover, audio deepfake can enhance social interactions and well-being of people with speaking disorders by providing a per

7 Varg Folkman, “Pakistan’s Imran Khan Uses AI to Make Victory Speech from Jail”, Politico, 11 February 2024.

8 Change the Ref, “UnfinishedVotes.com”, YouTube, 2 October 2020.

9 Dave Itzkoff, “The ‘South Park’ Guys Break Down Their Viral Deepfake Video”, The New York Times, 29 October 2020.

10 Dezeen, “Museum Creates Deepfake Salvador Dalí to Greet Visitors”, YouTube, 28 May 2019.

11 Vajira Thambawita et al., “DeepFake Electrocardiograms Using Generative Adversarial Networks Are the Beginning of the End for Privacy Issues in Medicine”, *Nature Scientific Reports*, vol. 11, Article number: 21896 (2021). Vera Sorin et al., “Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) - A Systematic Review”, *Academic Radiology*, vol. 27, Issue (August 2020). Bingquan Zhu et al., “Deepfakes for Medical Video De-Identification: Privacy Protection and Diagnostic Information Preservation”, *AIES ’20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020). Hoo-Chang Shin et al. “Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks”, *Simulation and Synthesis in Medical Imaging: Third International Workshop*, (2018).

sonalised synthetic voice, moving beyond the robotic voices used in traditional speech-assistive devices.[12]

In the context of justice and security, deepfake technology can offer useful applications. Its ability to create realistic and innovative facial reconstruction can aid forensic facial reconstruction. Similarly, deepfakes can reconstruct crime scenes to support criminal investigations. Finally, law enforcement authorities can use deepfake tools to identify and apprehend online child abusers.

“

This technology enables the smooth translation of auditory educational material across languages. Additionally, it can significantly lower the production costs of audiovisual educational content

Conclusion

Due to its misuses, deepfake technology understandably carries a negative connotation. Nevertheless, it is crucial to recognise that this technology is not solely a threat to human rights, but also an enabler for their realisation. In this light, deepfakes should be considered a dual-use technology, warranting a more nuanced and balanced public image that reflects its duality. Efforts to regulate this technology should likewise consider this duality and aim to minimise its misuses through proportionate measures while maximising its prosocial uses.

12 https://projectrevoice.org/.

Bibliography

Change the Ref, "UnfinishedVotes.com", YouTube.

Cole, Samantha (2017). "AI-assisted fake porn is here and we are all f**ked, Vice, 11 December.

Dezeen, "Museum creates deepfake Salvador Dalí to greet visitors", YouTube.

Finney Boylan, Jennifer (2018). "Will deep-fake technology destroy democracy?", The New York Times, 17 October.

Folkman, Varg (2024). "Pakistan's Imran Khan uses AI to make victory speech from jail", Politico, 11 February.

Godulla, Alexander, and others (2021). "Dealing with deepfakes - An interdisciplinary examination of the state of research and implications for communication studies", *Studies in Communication and Media*, Vol. 10 Issue 1, Accessed on 22 August 2025.

Itzkoff, Dave (2020). "The 'South Park' guys break down their viral deepfake video", The New York Times, 29 October.

Shin, Hoo-Chang, and others (2018). "Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks", *Simulation and Synthesis in Medical Imaging: Third International Workshop*.

Snow, Jackie (2017). "AI could set us back 100 years when it comes to how we consume news", MIT Technology Review, 7 November.

Sorin, Vera, and others (2020). "Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) - A Systematic Review", *Academic Radiology*, vol. 27, Issue 8.

Thambawita, Vajira, and others (2021). "DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine", *Nature Scientific Reports*, vol. 11, Article number: 21896.

The Supreme Court of the United States, *Ashcroft v. Free Speech Coalition*, 535 U.S. 234, April 16, 2002.

van Huijstee, Mariëtte, and others (2021). *Tackling deepfakes in European policy*, (Brussels, European Parliamentary Research Service).

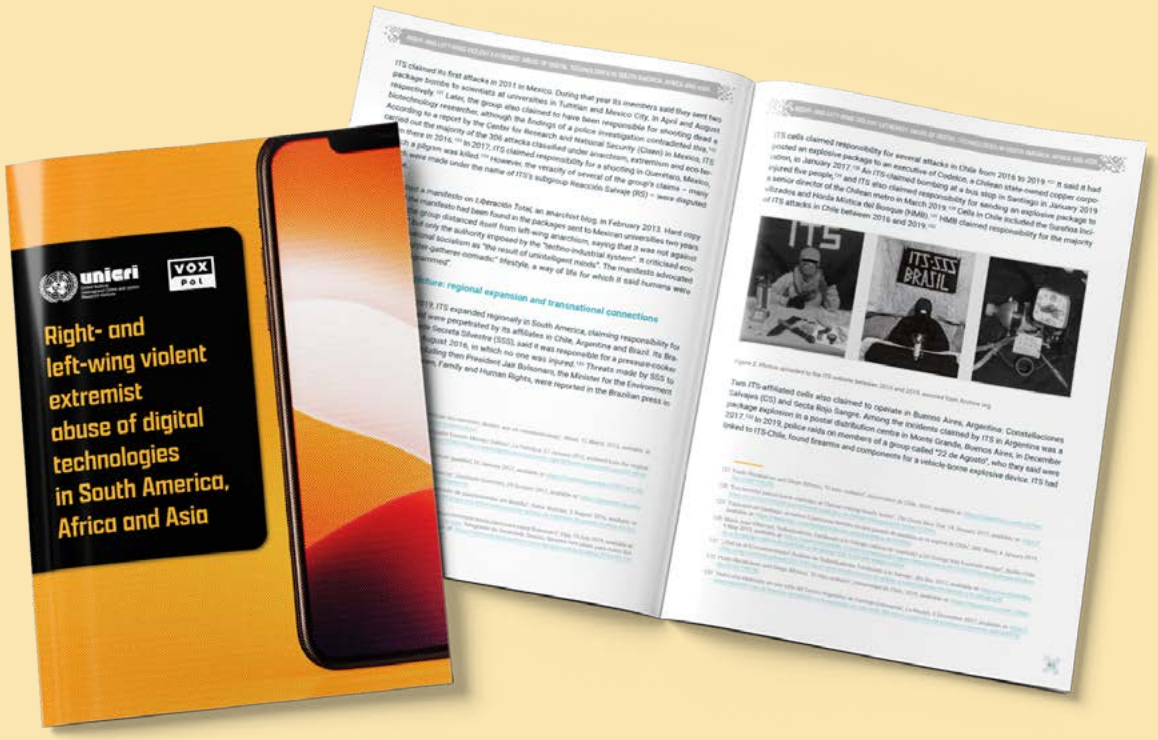
Yavuz, Can (2004). "A multidisciplinary look at history and future of deepfake with Gartner Hype Cycle", *IEEE Security & Privacy*, Vol. 2, Issue 2024.03.

Zhu, Bingquan, and others (2020). "Deepfakes for Medical Video De-Identification: Privacy Protection and Diagnostic Information Preservation", *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

ABOUT THE AUTHORS

Can Yavuz is a PhD student at Ghent University, Faculty of Law and Criminology.

Gert Vermeulen is Professor of International and European Criminal Law and Data Protection Law, and Director of the Institute for International Research on Criminal Policy, at Ghent University.



Download UNICRI publications

The exploitation of digital technologies by non-state armed groups in Colombia

by Arthur Bradley and Ottavia Galuzzi

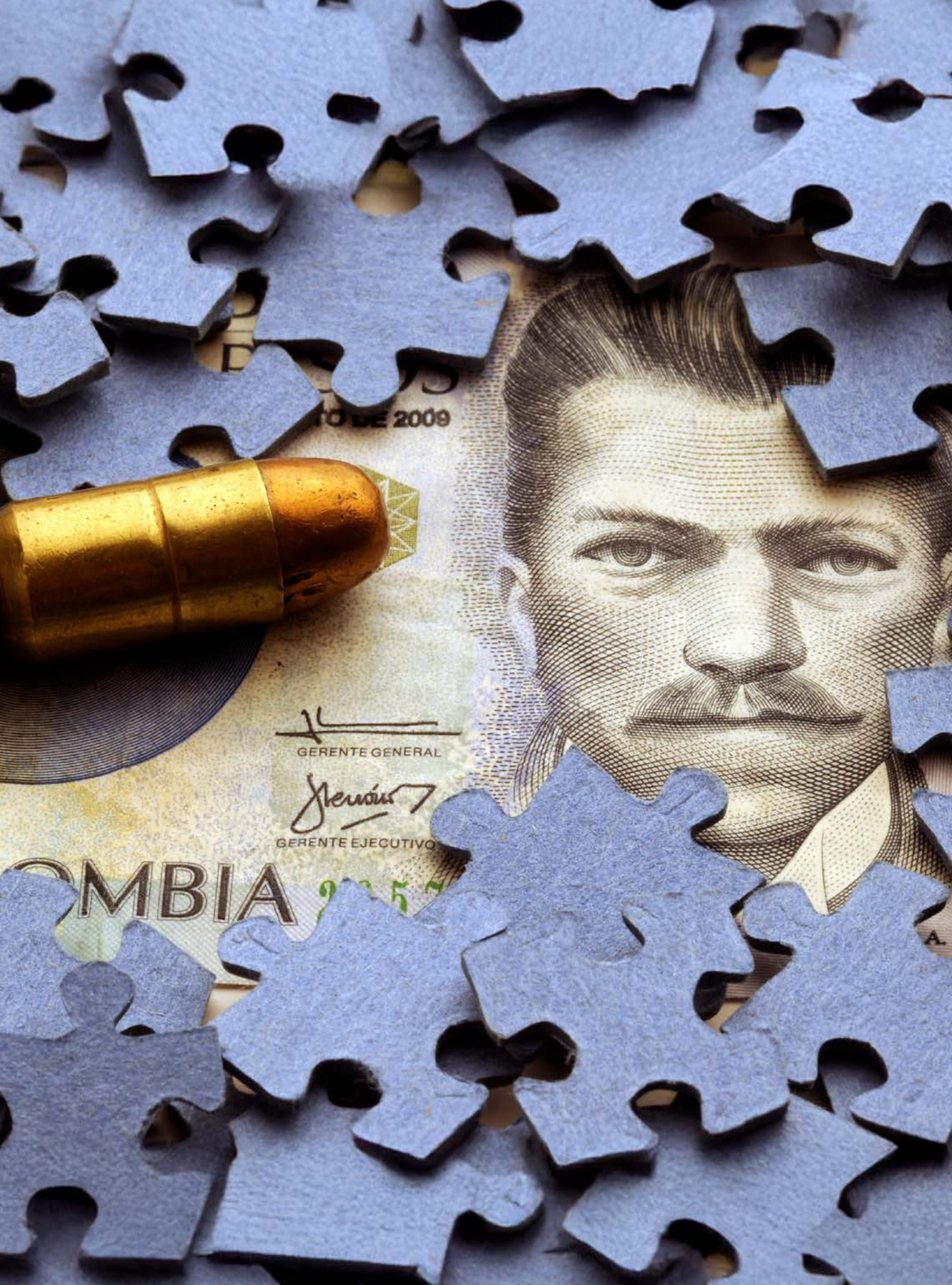
Armed non-state groups in Colombia have recently strengthened their presence through a series of attacks, the impact of which has been amplified by the widespread exploitation of digital technologies by these groups. On 21 August 2025, two separate attacks killed a total of 18 people and injured dozens of others in Colombia. In Amalfi, Antioquia, armed militants attacked a police helicopter carrying personnel overseeing the removal of coca crops, used in illicit drug production.[1] In the southwestern city of Cali, a vehicle loaded with explosives was detonated on a street near a military school. [2] Colombia's authorities attributed the Cali attack to the Estado Mayor Central, or Central General Staff (EMC), while the Ejército de Liberación Nacional (ELN), or National Liberation Army, claimed responsibility for the events in Amalfi, although the Government accused the Estado Mayor de los Bloques y Frentes (EMBF) of being responsible.[3]

Dissident factions of the FARC (FARC-D), ELN and other non-state armed groups have posed such threats to the Colombian state and its population for several decades. FARC-D denotes the dissident factions that rejected or later abandoned the 2016 peace accord, primarily split into the

1 "Two separate guerilla attacks kill 18 in Colombia", *France 24,* August 2025.

2 "Two separate guerilla attacks kill 18 in Colombia", *France 24,* August 2025.

3 United Nations Security Council, "United Nations Verification Mission in Colombia – Report of the Secretary General", September 2025.


GERENTE GENERAL
GERENTE EJECUTIVO

EMC and Segunda Marquetalia (SM); some affiliated accounts continue to use FARC-EP branding. Despite ongoing peace negotiations in the country, groups such as ELN and FARC-D are increasingly exploiting digital platforms such as TikTok and static websites to communicate externally and recruit, particularly minors.

A study published by the Colombian Organized Crime Observatory in 2022 identified 1,020 cases of child recruitment by guerrilla groups between 2017 and 2020 in all but two of the country's departments. In the first half of 2024, according to government data, 159 children were enlisted by armed groups, and 55 additional recruitment cases were recorded during the first half of 2025.[4] However, the true figures of child recruitment are likely to be significantly higher, as many families refuse to report their missing children for fear of reprisal.[5]

While child recruitment is not a new issue faced by the Government of Colombia, militant groups' increasing use of social media for this purpose has made it more difficult to counter.[6]

Abuse of digital technologies

As part of UNICRI's workstream on Cybercrime and Online Harms, UNICRI conducted research into the increasing exploitation of social media and other digital technologies by non-state armed groups in Colombia, which in recent years has coincided with a broader increase in Internet access among the general population.[7]Communications by non-state armed groups and their members have progressively shifted from paper pamphlets to digital messaging, including in private encrypted spaces such as WhatsApp and Telegram,[8] as well as in more public digital spaces like TikTok, X, Instagram, YouTube, Facebook and Russia-based video platforms.

"A study published by the Colombian Organized Crime Observatory in 2022 identified 1,020 cases of child recruitment by guerrilla groups

The most powerful armed groups, including the ELN and the EMC, maintain an official presence on mainstream social media platforms. Broadly, the groups use these channels to share official statements, often political rather than militant in nature, to position themselves in the national debate. Messaging shared here typically aims to present the organizations as legitimate political actors, rather than violent criminals, while on platforms with younger audiences, the videos openly flaunt a narco lifestyle.[9]

The BBC reported in June 2024 that the EMC maintained a WhatsApp group with journalists, in addition to its profiles on mainstream social media.[10] There are signs that international technology companies are moderating accounts operated by non-state armed groups in Colombia,[11] although many of these accounts still remain active. Groups routinely repost identical videos across multiple hosts to outlast takedowns, while cycling handles and logos to evade automated filters.

4 Rachelle Krygier and Laura García, "The school children being lured by rebels on TikTok", BBC News, June 2024; Defensoria del Pueblo, "Reclutamiento de niñas, niños y adolescentes en Colombia durante el primer semestre de 2025", July 2025.

5 Elizabeth Dickinson, "Colombia's Stolen Children. Bogotá Must Do More to Stop Armed Groups From Recruiting Minors", Foreign Affairs, May 2025.

6 Rachelle Krygier and Laura García, "The schoolchildren being lured by rebels on TikTok", BBC News, June 2024.

7 Interview with a regionally based analyst, 2024.

8 Interview with Camilo Tamayo Gomez, University of Huddersfield, 2024.

9 Interview with a regionally based analyst, 2024.

10 Rachelle Krygier and Laura García, "The schoolchildren being lured by rebels on TikTok".

11 Luis Jamie Acosta, "Social networks clamp down on Colombian FARC dissident accounts", Reuters, January 2021.

The screengrab (Figure 1) shows a tweet posted by one of a network of accounts affiliated with the EMC's FARC-EP (EP – People's Army) on X, including video footage of armed individuals from the Jaime Martínez Front planting trees as part of what it describes as "reforestation day". The accompanying text criticizes the then upcoming United Nations Biodiversity Conference in Colombia in October 2024 as an event that it claims is "disguised as environmentalism" but "promotes militarism". Some official FARC-EP accounts identified during this research were removed by X during the analysis phase, including accounts that had been active since as early as 2022.

Figure 1.
An account on X affiliated with a dissident FARC group, captured in July 2024

WHOIS information (the standard public record of domain ownership) for a website likely affiliated with FARC-SM shows that it was created in October 2019 and is registered with a domain registrar based in the United States. Archived versions of the site suggest that it has been active for most of the five years since it was created. Open-source research indicates that FARC-SM also maintains at least three channels on Telegram and an official account on X. In addition, the group has posted videos on multiple platforms simultaneously – a tactic also used by other threat actors to maximize the digital lifespan of content in the face of removal by individual technology companies.[12]

12 Stuart Macdonald and Sean McCafferty, "Online Jihadist Propaganda Dissemination Strategies", *VOX-Pol*, March 2024.

In May 2024, for example, a video featuring Iván Márquez, FARC-SM leader widely reported as killed, though this remains unconfirmed, was shared on a FARC-linked website with links to copies on Terabox, Proton Drive, Google Drive, and WeTransfer.

The ELN has a more sophisticated and wide-reaching public online presence than the EMC or other Colombian-based groups.[13] This analytical research identified accounts affiliated with the ELN on multiple platforms, including X, Mastodon, Instagram, Facebook, Telegram, YouTube, Internet Archive, and a network of ten websites, including those dedicated to its provincial fronts. Its radio station, Antorcha Stereo,[14] advertises affiliated pages on TikTok, Facebook, Instagram and X via its own page on Beacons.ai, an "all-in-one creator platform" designed for social media influencers.[15]

In addition to digital messaging curated online by the leadership of these non-state armed groups, their individual members also have an increasing presence on digital platforms. Multiple recent studies have found accounts run by members or supporters of dissident FARC groups, including on Facebook and TikTok.[16] It is likely that such accounts exist without leadership approval. In June 2024, a member of the EMC's Dialogue Commission told the press that the group was "trying to control" the use of social media by its members, given the associated security risks of public photos or other content, including innocuous backgrounds and image patterns, that might give away the identity and location of fighters.[17]

Searches for related keywords and hashtags on TikTok reveal multiple accounts affiliated with dissident FARC groups.

13 Interview with Camilo Tamayo Gomez, University of Huddersfield, 2024.

14 Daniel Pardo, "Antorcha Stereo, la polémica emisora de la guerrilla colombiana que se escucha en Venezuela", *BBC News Mundo*, October 2015.

15 https://beacons.ai/i/about-beacons.

16 "Colombia's guerrilla recruitment video problem", BBC Trending Podcast, June 2024.

17 Rachelle Krygier and Laura García, "The schoolchildren being lured by rebels on TikTok".

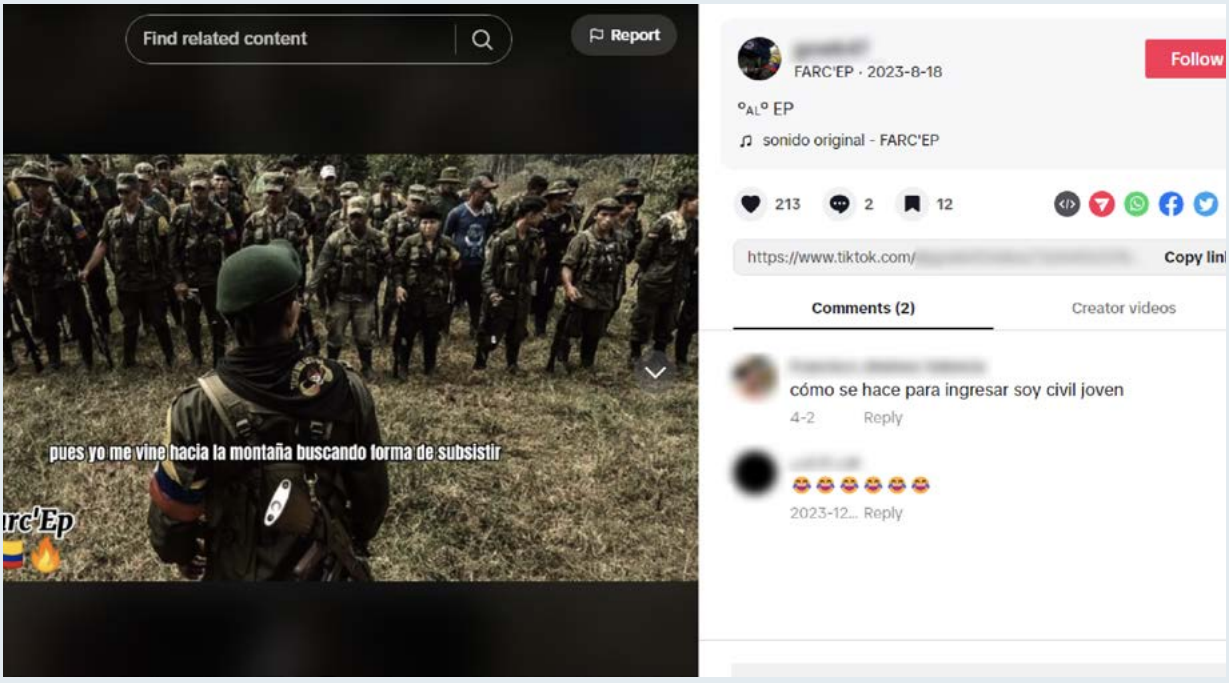


Figure 2. A TikTok account affiliated with FARC-EP, captured in July 2024.

The accounts are identifiable by their use of FARC logos and particular emojis, including the green leaf to signify the coca plant, often combined with the ninja emoji and the Colombian flag. These accounts typically promote the narco lifestyle[18] with images of firearms and fighters in the jungle. Posts are often accompanied by music associated with armed groups. Interactions in the comments with non-members, some of whom express interest in joining, suggest that such content serves as a recruitment tool on TikTok.

> **"In addition to digital messaging curated online by the leadership of these non-state armed groups, their individual members also have an increasing presence on digital platforms**

There are a few indications that non-state armed groups in Colombia have the capability to exploit digital technologies for more offensive technical cyber operations. However, militants are reported to routinely harass civilians online, in particular those they allege to be government informants or allied with rival armed groups.[19] Experts interviewed as part of this brief research believed that non-state armed groups there are likely interested in expanding their cyber portfolio.[20] It is probable that higher-ranking criminals within

18 Elizabeth Dickinson, "Colombia's Stolen Children. Bogotá Must Do More to Stop Armed Groups From Recruiting Minors", Foreign Affairs, May 2025.
19 Interview with a regionally based analyst, 2024.
20 Interview with a regionally based analyst, 2024.

these groups are exploiting encrypted networks and cryptocurrencies, access to which can be hampered by poor Internet connectivity in rural areas. This assumption is supported by the growing use of Starlink Internet in regions with minimal state presence but significant activity by armed groups, such as remote areas of the Amazon.[21]

The use of the Internet by non-state armed groups in Colombia is indicative of broader efforts by violent actors to exploit digital technologies to further their objectives, including via propaganda, recruitment, internal communication, and financing. Such issues require cross-sector responses involving stakeholders from both the public and private sectors, including collaborative working relationships between governments and the technology sector. This is particularly important in Colombia and other countries in the Global South, where existing initiatives must be leveraged to address the challenges posed by militancy in the digital realm. The use of digital technologies by non-state armed groups, as by the general population, is an integral part of day-to-day communication and productivity, which means that neither online nor offline spaces can be studied in isolation from one another.

“

Groups routinely repost identical videos across multiple hosts to outlast takedowns, while cycling handles and logos to evade automated filters

ABOUT THE AUTHORS

Arthur Bradley is an specialist in tracking and analyzing terrorist and other malevolent use of online platforms. He provides consultancy services to the private sector, non-governmental organizations, IGOs, academia, public sector institutions, and law enforcement agencies. Prior to working independently, Arthur was the OSINT manager at Tech Against Terrorism.

Ottavia Galuzzi is associate expert, where she works on projects aimed at preventing and countering cybercrime and online harms. She has experience working in the private, public and civil society sectors as a cybersecurity, intelligence and P/CVE consultant.

21 Interview with a regionally based analyst, 2024; "Nicolás Maduro Moros and 14 Current and Former Venezuelan Officials Charged with Narco-Terrorism, Corruption, Drug Trafficking and Other Criminal Charges", *US Department of Justice Office of Public Affairs Press Release,* March 2020.

unicri
United Nations Interregional Crime and Justice Research Institute

University for Peace

MASTER OF LAWS LLM

IN CYBERCRIME, CYBERSECURITY AND INTERNATIONAL LAW

LL.M. Office

E-mail: unicri.llmcyber@un.org

Tel.: (+39) 011 6537 157-154

(+39) 011 6537 111

Postal Address:

LL.M. in Cybercrime, Cybersecurity and International Law

UNICRI - Viale Maestri del Lavoro, 10

10127 - Torino, ITALY

unicri
United Nations Interregional Crime and Justice Research Institute

University for Peace

LL.M. Office

E-mail: unicri.llm@un.org

Tel.: (+39) 011 6537 157-154

(+39) 011 6537 111

Postal Address:

LL.M. in Transnational Crime and Justice

UNICRI - Viale Maestri del Lavoro, 10

10127 - Torino, ITALY

MASTER OF LAWS LLM

IN TRANSNATIONAL CRIME AND JUSTICE

Explore our Master of Laws (LL.M.) program at **www.unicri.org** and take the next step in your legal education.

The dual role of satellite internet in fragile contexts: opportunities for development and challenges for security

by Gaetano Sicolo

Today's global society has become increasingly reliant on digital infrastructure and connectivity. The growth of technology and the ability to stay connected offer an opportunity for low- and middle-income countries (LMICs) to close the digital gap and accelerate their development. One key aspect of this change is the rapid rise of satellite Internet, which offers fast and affordable connectivity in unserved and underserved communities.[1]

"The growth of technology and the ability to stay connected offer an opportunity for low- and middle-income countries (LMICs) to close the digital gap and accelerate their development

While it may sound like science fiction, this same technology has been exploited by terrorist and violent extremist groups operating in fragile regions to facilitate their operations. It has been demonstrated that such groups use satellite-based communications to coordinate operations in remote areas and spread propaganda in places that were previously inaccessible.[2]

The role of satellite networks in developing contexts

Digital technologies and Internet connectivity can drive development, especially in LMICs, as being connected provides better access to information and increases communication, along with new opportunities for learning and work. While this appears positive, the impact of these benefits depends on factors such as infrastructure availability, service affordability, access to devices, connection security, and the population's level of digital literacy skills.[3]

These positive factors led the United Nations to prioritize universal connectivity as a goal to achieve by 2030,[4] with satellite connectivity having the potential to increase both fixed and mobile broadband access in areas where fiber Internet is scarce. Economically, studies show that a 10%

1 International Telecommunication Union, "Managing Spectrum for Evolving Technologies", *ITU News Magazine*, No. 5, 2019.
2 Global Initiative Against Transnational Organized Crime, "Risk Bulletin of Illicit Economies in West Africa", Issue 12, May 2025.
3 International Telecommunication Union, "Achieving Universal and Meaningful Digital Connectivity: Setting a Baseline and Targets for 2030", 2021.
4 United Nations, "Roadmap for Digital Cooperation", Report of the Secretary-General, 2020.

rise in Internet penetration can boost gross domestic product (GDP) per capita growth by 2% to 2.3% for fixed broadband and 2.5% to 2.8% for mobile broadband in developing countries.[5]

Economic growth, in turn, leads to improvements in education and healthcare. Higher-quality education boosts literacy rates and, consequently, digital literacy, addressing one of the main barriers to Internet use along with affordability.[6] At the same time, satellite Internet has been used for telemedicine and remote diagnostics in rural areas that lack resident doctors or sufficient medical infrastructure, providing a solution for countries where building terrestrial infrastructure is challenging.[7]

> **"In the Sahel and wider Sub-Saharan Africa, criminals not only exploit Starlink terminals but also smuggle them across the region into countries where it is still illegal**

Although satellite connectivity has the potential to provide both universal and meaningful coverage, the same features that make satellite networks advantageous also introduce vulnerabilities.

In fragile regions, terrorist and violent extremist groups can exploit these technologies to coordinate operations and spread propaganda, illustrating the double-edged nature of this emerging technological infrastructure.[8]

Security risks and misuse by violent extremist organizations

When considering the nexus between transnational organized crime and terrorism, new technologies should be

5 International Telecommunication Union, "Economic Impact of Broadband in LDCs, LLDCs and SIDS: An Empirical Study", 2019.

6 International Telecommunication Union, "ICTs, LDCs and the SDGs: Achieving Universal and Affordable Internet in the Least Developed Countries", 2018.

7 Eutelsat, "Telehealth for Remote Communities".

8 Gaetano Sicolo, "Dark Signals: The Growing Threat of Satellite Internet in Extremist Networks", *Global Network on Extremism & Technology,* 2024.

viewed within a broader context. Terrorists and violent extremists have been increasingly using digital platforms for two main reasons: to promote extremist content and to coordinate agendas.[9]

In the past, groups like the Islamic State West Africa Province (ISWAP) relied on slow, high-latency satellite Internet to communicate with other extremist organizations in the region.[10] For their needs, Thuraya Wi-Fi offered Internet access at 60 Kbps within a 30-metre range, but that changed with the arrival of Starlink. Starlink uses Low-Earth Orbit (LEO) satellites, allowing higher download speeds of around 200 Mbps in areas lacking stable terrestrial networks.[11]

In the Sahel and wider Sub-Saharan Africa, criminals not only exploit Starlink terminals but also smuggle them across the region into countries where it is still illegal.[12] Unfortunately, this offers numerous advantages to these threat actors, enabling them to gain an edge over national forces and use instant messaging applications to spread propaganda or recruit new fighters.

Looking ahead, satellites may become increasingly targeted by cyber-attacks by violent extremists as their technological capabilities advanc- Currently, satellites perform essential military functions, and many global services rely heavily on this infrastructure. This dual-use nature indicates that a successful attack could have far-reaching consequences beyond vulnerable regions. It also underscores the importance of viewing space-based communication networks as critical infrastructure, since their disruption could have widespread impacts.[13]

Satellite networks as critical infrastructure

Identifying infrastructure as critical allows governments to recognize and address its vulnerabilities, and satellite networks are among the most exposed to malicious interfer-

9 Folahanmi Aina et al., "The "Webification" of Jihadism: Trends in the Use of Online Platforms, Before and After Attacks by Violent Extremists in Nigeria", *Global Network on Extremism & Technology*, 2023.

10 Malik Samuel, "ISWAP's Use of Tech Could Prolong Lake Chad Basin Violence", *Institute for Security Studies,* 2023.

11 Abdoulaye Mamane, 'Signature d'accord sur l'internet haut débit: Global licensing and activation I Starlink s'engage à fournir l'internet haut débit au Niger', *Le Sahel,* 2024.

12 Global Initiative Against Transnational Organized Crime, "The Shadow Constellation: How Starlink Devices Are Shaping Conflict and Crime in the Sahel", *Observatory of Illicit Economies in West Africa,* Issue 12, 2025.

13 Meghan Bartels, "Why Satellites Need Cybersecurity Just Like You", *Space.com,* 2018.

ence. Radiofrequency attacks such as spoofing are already widespread and require limited technical expertise.[14] Recent civil aviation data illustrate the scale of the threat: on average, 1,500 flights per day were affected by spoofing in 2024, with more than 41,000 incidents recorded between July and August alone.[15]

Satellites and the broader space ecosystem should be regarded as critical infrastructure, precisely because they support national security, public health, economic activity, and the functioning of modern societies. In practice, they are connected to nearly every vital sector, and studies have indicated that a successful cyber-attack could disrupt these sectors within just 12 hours.[16] This risk was highlighted during the U.S. *Hack-A-Sat* competition, where ethical hackers demonstrated how vulnerabilities in satellite systems could be exploited if not properly secured, emphasising the urgency of adopting stronger security standards to prevent malicious actors from exploiting the same weaknesses.[17]

> **Identifying infrastructure as critical allows governments to recognize and address its vulnerabilities, and satellite networks are among the most exposed to malicious interference**

The current approach to satellite security remains fragmented. Some national strategies classify space assets as part of critical infrastructure, but international coordination remains limited.[18] New constellations are being launched faster than security frameworks can adapt, resulting in significantly higher costs for responding to major incidents, especially in sectors that rely heavily on satellite networks.

As satellite constellations continue to grow and more services move to space-based networks, the question is no longer whether they should be protected but how. This urgency calls for coordinated strategies to safeguard satellites while enabling their development potential. The next step is to consider what those strategies might look like.

Conclusion and recommendations

The momentum generated by satellite infrastructure needs careful management, considering both its growth potential and risks. As mentioned earlier, its disruption could trigger a chain reaction across critical sectors, from communications and transportation to public health and emergency services. Given the importance of these systems, governments cannot protect them alone; instead, a collaborative approach involving international organizations, private operators, and Member States would be more effective.

While coordination mechanisms already exist under the International Telecommunication Union and the Committee on the Peaceful Uses of Outer Space (COPUOS), they should be strengthened and better aligned, moving from voluntary reporting[19] to mandatory public-private threat responses. Ensuring the proper

14 Nicolò Boschetti et al., "Commercial Space Risk Framework Assessing the Satellite Ground Station Security Landscape for NATO in the Arctic and High North", *IEEE*, 2022.

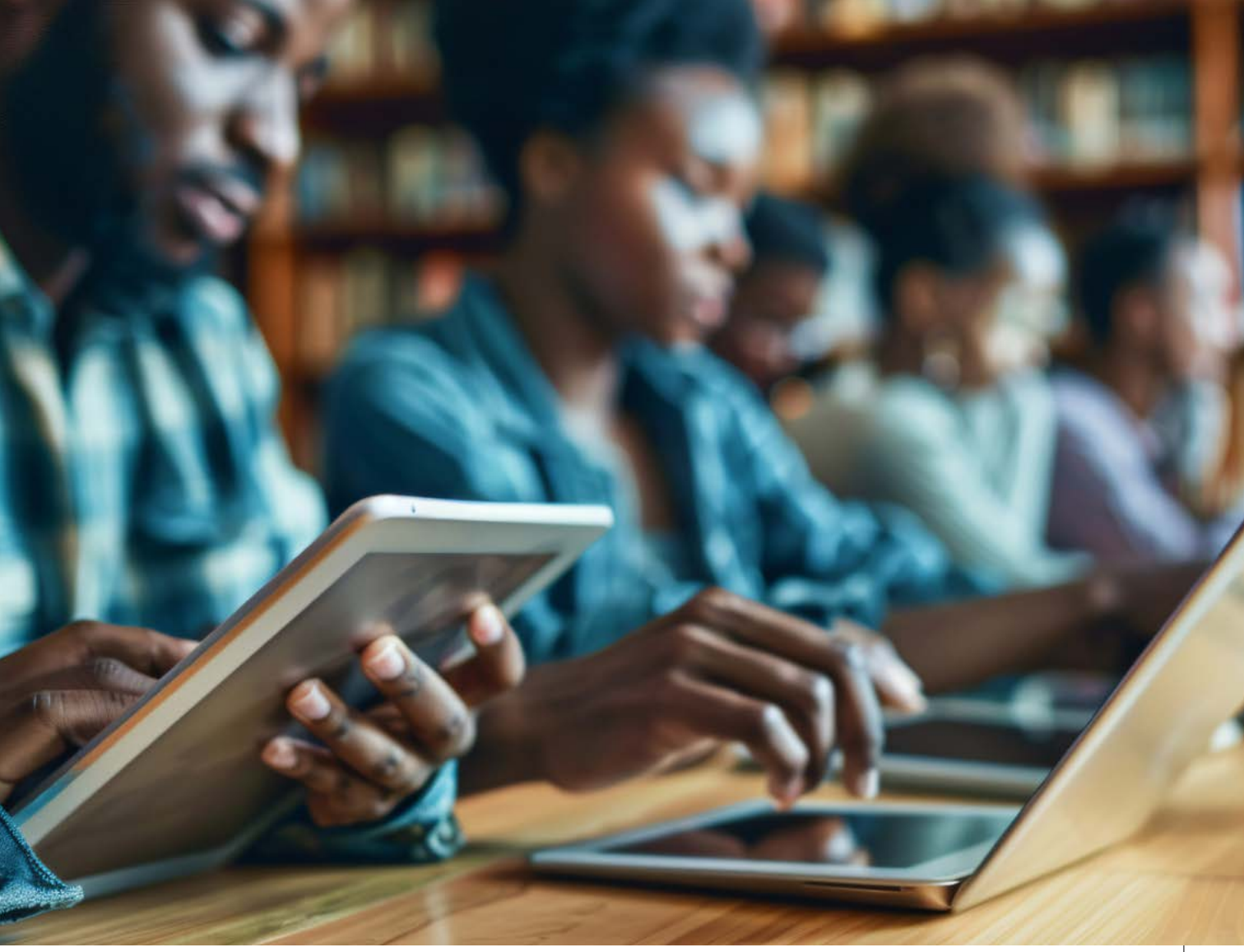
15 Ops Group, "Final Report of the GPS Spoofing workgroup", 2024.

16 Linda Dawson, "Life Without Satellites in War in Space", *Springer*, 2018.

17 Brett Tingley, "These 3 Teams Just Hacked a US Air Force Satellite in Space... and Won Big Cash Prizes", *Space.com*, 2023.

18 Rian Davis et al., "Space as Critical Infrastructure: An In-Depth Analysis of U.S. and EU Approaches", *Acta Astronautica*, vol. 225, 2024.

19 United Nations Office for Outer Space Affairs, "Guidelines for the Long-term Sustainability of Outer Space Activities of the Committee on the Peaceful Uses of Outer Space", 2021.



operation of satellite networks should be a top priority, and security strategies should focus on making sure that the benefits are not compromised by their vulnerabilities. As developing regions expand their Internet access, much of their growth will depend on the safety and affordability of space-based communication systems. Therefore, strengthening cooperation among Member States, international organizations, and private companies will be crucial to bridging digital gaps and ensuring that satellites are consistently protected as critical infrastructure.

> **“As satellite constellations continue to grow and more services move to space-based networks, the question is no longer whether they should be protected but how**

ABOUT THE AUTHOR

Gaetano Sicolo is a graduate student in Area and Global Studies for International Cooperation at the University of Turin, with research experience in geopolitics, technology, and security. He has contributed to the work of the Global Network on Extremism and Technology and the Institute for the Analysis of International Relations (Istituto Analisi Relazioni Inter nazionali), producing analyses on the implications of digital connectivity, satellite internet, and violent extremism in Sub-Saharan Africa.

UNSCR 1540 imposes binding obligations on States to prevent non-State actors from developing, acquiring, manufacturing, possessing, transporting, transferring or using nuclear, chemical and biological weapons and their means of delivery, and to establish effective domestic controls to that end

New technologies and the manipulation of facts in armed conflicts

by Ari Koutale Tila Boulama Mamadou

Framing context

Armed conflict is undergoing a structural transformation driven by rapid technological innovation. Artificial intelligence (AI), autonomous and semi-autonomous weapon systems, algorithmic decision-support tools, cyber operations and digital platforms for information manipulation are increasingly embedded in doctrine and battlefield practice.[1] This transformation does not only change the "means and methods" of warfare; it reshapes how facts are produced, how responsibility is attributed, and how international humanitarian law (IHL) is interpreted and enforced.[2]

IHL is built on a factual architecture: who was targeted, what was known (or reasonably believed) at the time, which precautions were taken, and whether expected civilian harm was excessive in relation to the anticipated military advantage. In conventional settings, those judgements presume human contextual reasoning and a traceable chain of decision-making. In today's conflicts, however, technical systems increasingly mediate both decision and documentation. Algorithmic recommendations can shape targeting decisions; higher autonomy can compress the time available for legal review; and digital manipulation can undermine the reliability of information used by fact-finders and courts.[3]

These dynamics intersect with the preventive logic of United Nations Security Council resolution 1540 (2004). UNSCR 1540 imposes binding obliga-

1 International Committee of the Red Cross (ICRC), International Humanitarian Law and the Challenges of Contemporary Armed Conflicts, 2024.

2 Ibid.

3 Office of the United Nations High Commissioner for Human Rights (OHCHR), Berkeley Protocol on Digital Open Source Investigations, last accessed December 18, 2025.

“The humanitarian cost is borne by people on the ground: civilians who look “abnormal” to an algorithm, humanitarian staff moving near contested areas, and communities whose daily routines are misread as hostile patterns

tions on States to prevent non-State actors from developing, acquiring, manufacturing, possessing, transporting, transferring or using nuclear, chemical and biological weapons and their means of delivery, and to establish effective domestic controls to that end.[4]

This article argues that emerging military technologies generate a dual risk. First, they strain compliance with IHL principles such as distinction, proportionality and precaution by shifting critical judgement toward machine-mediated processes. Second, they destabilise the evidentiary foundations needed to investigate violations, attribute responsibility, and support preventive action under frameworks such as UNSCR 1540.

Autonomous and algorithmic targeting: reconfiguring distinction, proportionality and human responsibility

The targeting cycle is a central pressure point. AI systems are increasingly used to fuse intelligence, prioritise targets, support battle-damage assessment, and accelerate operational tempo. In parallel, some weapon platforms incorporate higher degrees of autonomy. The International Committee of the Red Cross (ICRC) describes autonomous weapon systems as those that can select and apply force to targets without human intervention after activation, based on sensor information and a generalised "target profile", meaning the user may not choose or even know the specific target(s) at the moment force is applied.[5]

Distinction becomes more difficult when classification is statistical rather than contextual. Under IHL, attackers must distinguish between civilians and combatants, and between civilian objects and military objectives. Human decision-makers interpret context and uncertainty, including patterns of civilian life and the presence of protected objects.
Machine-learning systems, by contrast, infer categories from training data. In complex environments, irregular forces, civilians intermingled with fighters, degraded visibility, adversarial deception systems can misclassify in ways that are operationally subtle but legally decisive. The humanitarian cost is borne by people on the ground: civilians who look "abnormal" to an algorithm, humanitarian staff moving near contested areas, and communities whose daily routines are misread as hostile patterns.

> **"Accountability depends on the ability to establish credible facts, yet contemporary conflicts increasingly unfold in an "epistemic battlespace" where information is contested and deliberately manipulated**

Proportionality and precaution are even less amenable to automation. Proportionality prohibits attacks expected to cause incidental civilian harm that would be excessive in relation to the concrete and direct military advantage anticipated. Even if algorithms can estimate blast effects or probabilities, proportionality still demands a qualitative legal judgement about what is "excessive" in context, based on information reasonably available at the time. Likewise, precaution

4 UN Security Council, Resolution 1540 (2004), S/RES/1540 (adopted on 28 April 2004), UN Digital Library record, last accessed December 18, 2025.
5 ICRC, "ICRC position on autonomous weapon systems" definition of AWS and "target profile" language, last accessed 17 December 2025.

requires all feasible steps to minimise harm and to cancel or suspend attacks when circumstances change.[6] High-tempo machine-assisted targeting can narrow the "reconsideration window", the very space in which human judgement is supposed to correct error, uncertainty, or new civilian presence. A core legal risk is quiet norm-drift: allowing the technical limits of a system to redefine what counts as a "feasible" precaution.

Accountability is where the IHL and 1540 lenses converge. Technical systems can diffuse agency across operators, commanders, developers, contractors, data suppliers, and procurement authorities; opacity can frustrate after-action review. Yet IHL and international criminal law remain anchored in human responsibility and traceable decision-making. Meaningful human control is therefore not a purely ethical add-on; it is a practical way of preserving a chain of responsibility, ensuring that a legally accountable human can understand the basis for action, question it, and stop it when legal doubt arises. The ICRC has explicitly called for regulation of autonomous weapon systems and for preserving human control over the use of force.[7]

From a UNSCR 1540 perspective, traceability also supports prevention. Many enabling components of algorithmic targeting high-resolution sensors, geospatial analytics, communications equipment, drone subsystems, and certain software stacks are dual-use and increasingly accessible through commercial markets. Weak controls can enable diversion and capability transfer to non-State actors. Strengthening review, documentation, and human control is therefore simultaneously a humanitarian safeguard and a preventive measure consistent with 1540 obligations.[8]

Manipulation of facts and the crisis of digital evidence in armed conflict

If targeting is one pressure point, evidence is the other. Accountability depends on the

6 International Humanitarian Law and the Challenges of Contemporary Armed Conflicts: Building a Culture of Compliance for IHL to Protect Humanity in Today's and Future Conflicts, Report to the 34th International Conference of the Red Cross and Red Crescent (Power of Humanity), Geneva, 28–31 October 2024, September 2024, last accessed 17 December 2025.

7 Office of the United Nations High Commissioner for Human Rights (OHCHR), Berkeley Protocol on Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source Information in Investigating Violations of International Criminal, Human Rights and Humanitarian Law, 2022, last accessed 18 December 2025.

8 UN Security Council, Resolution 1540 (2004), S/RES/1540 (adopted 28 April 2004, last accessed 18 December 2025.

ability to establish credible facts, yet contemporary conflicts increasingly unfold in an "epistemic battlespace" where information is contested and deliberately manipulated. Synthetic media, including deepfakes and AI-generated audio, can fabricate incidents, misattribute attacks, or discredit legitimate reporting.[9] The strategic effect is corrosive: even authentic material becomes suspect, witnesses become easier to discredit, and truth becomes harder to prove in legal proceedings.

Computational propaganda adds scale and persistence. Automated accounts, bot networks and targeted disinformation can amplify false narratives, intimidate witnesses, and undermine trust in humanitarian organisations and investigators. The ICRC's Global Advisory Board report on digital threats frames "harmful information" as a key vector of civilian harm and a challenge for protection during armed conflict.[10]

Cyber operations intensify evidentiary vulnerability in a different way: by attacking the infrastructure that stores and transmits proof. Surveillance logs, metadata, communications records and weapon-system telemetry may be deleted, corrupted, encrypted or altered. When audit trails are erased, accountability can become practically "unprovable", not because harm did not occur, but because the evidentiary chain has been compromised.

In this context, "protecting civilians" includes protecting the informational conditions under which protection and accountability can function. When truth can be manufactured or erased at scale, both humanitarian and preventive security frameworks become harder to operationalise.

Digital evidence, verification and legal admissibility in international practice

International justice institutions and fact-finding mechanisms are adapting to the digital turn, but verification and admissibility remain fragile. Digital evidence raises persistent challenges: authentication (is it what it claims to be?), provenance (who handled it and how?), and preservation (was it altered, compressed, or stripped of metadata?). These are not merely technical questions; they affect fairness, reliability, and the probability that perpetrators are held responsible. The ICC's "Unified Technical Protocol (eCourt Protocol)" reflects the Court's move toward standardised electronic handling and submission of evidence and related materials.[11]

The Berkeley Protocol on Digital Open Source Investigations developed through OHCHR responds to these challenges by offering international standards and guidance on identifying, collecting, preserving, verifying and analysing digital open-source information for international criminal, humanitarian and human rights investigations.[12] Its core contribution is methodological consistency: a shared professional baseline that helps judges, commissions, and investigators evaluate how evidence was obtained and tested.

For investigators and humanitarian actors, verification is concrete work: preserving originals, capturing metadata, geolocating images, cross-checking sources, documenting provenance, and securing storage.

9 OHCHR, Berkeley Protocol on Digital Open Source Investigations, 2022.

10 Protecting Civilians Against Digital Threats During Armed Conflict : Final report of the ICRC's Global Advisory Board on Digital Threats During Armed Conflicts, 19 October 2023.

11 International Criminal Court (ICC), Unified Technical Protocol ("E court Protocol") for the Provision of Evidence, Witness and Victims Information in Electronic Form.

12 OHCHR, Berkeley Protocol on Digital Open Source Investigations, 2022.

Where resources are limited, uneven capacity produces uneven evidentiary quality and, ultimately, uneven access to justice. The ICRC has also highlighted that digital threats target humanitarian organisations themselves, making cybersecurity and information integrity part of protection and operational security.[13]

> **When audit trails are erased, accountability can become practically "unprovable", not because harm did not occur, but because the evidentiary chain has been compromised**

States can support evidentiary resilience by investing in digital forensics, standardising domestic procedures for receiving and protecting digital evidence, and enabling secure cooperation with civil society and technology companies that may hold relevant data. For UNSCR 1540, these investments align with national implementation measures that strengthen investigative capacity and improve prevention and enforcement against non-State actors.[14]

Toward an integrated normative framework: IHL, AI governance and UNSCR 1540

Technological change does not diminish legal responsibility; it increases the need for adaptive governance. The ICRC's 2024 "Challenges Report" underlines how contemporary conflicts raise evolving issues for IHL, including those linked to new technologies.[15] Building on this, a coherent response requires connecting four priorities.

Modernise weapons and methods review. Where Article 36 review obligations apply, they should address AI-enabled systems through realistic testing, evaluation of failure modes, and scrutiny of human–machine interaction, including how systems behave under uncertainty and in civilian-dense environments.[16] Reviews should explicitly consider evidentiary consequences: whether the system's operation will be reconstructible after the fact.

Operationalise meaningful human control. Human control must be substantive: informed operators, real intervention capacity, and command responsibility that does not defer to machine outputs. This is essential to preserve IHL's structure of accountability and civilian protection.[17]

Design for accountability and evidentiary resilience. Where possible, systems and processes should generate auditable records (logs, secure storage, documentation of targeting decisions and data inputs), while protecting sensitive information. These measures support IHL investigations and strengthen States' capacity to detect misuse, diversion, and illicit technical modification.

Align preventive controls with UNSCR 1540 realities. UNSCR 1540 requires States to adopt and enforce effective measures to prevent proliferation to non-State actors,

13 ICRC, Protecting Civilians Against Digital Threats During Armed Conflict, 2023.

14 UN Security Council, Resolution 1540 (2004), S/RES/1540.

15 ICRC, International Humanitarian Law and the Challenges of Contemporary Armed Conflicts (2024) sections addressing evolving challenges/new technologies and IHL.

16 International Humanitarian Law and the Challenges of Contemporary Armed Conflicts Building a Culture of Compliance for IHL to Protect Humanity in Today's and Future Conflicts, September 2024

17 ICRC Position on Autonomous Weapon Systems, Geneva, 12 May 2021.

including appropriate domestic controls.[18] In an era of AI- and cyber-enabled conflict, this implies expanding national risk assessments to include software and data-driven tools; strengthening export, brokering and end-use controls for relevant dual-use technologies; and developing structured cooperation with the private sector, which often holds key technical knowledge and data relevant to prevention.

If law is to remain credible, it must remain operational. The objective is not to "ban technology", but to ensure that innovation does not erode the legal duties that protect civilians and enable responsibility and prevention.

ABOUT THE AUTHOR

Ari Koutale Tila Boulama Mamadou is a diplomat, First Secretary at the Permanent Mission of Niger to the United Nations in New York, and focal point for Niger to the Security Council›s 1540 Committee. He is in charge of disarmament and international security, including counter-terrorism. He also handles legal issues and ICT-related matters. In the academic field, he holds a Master's degree in International Relations, specializing in Security Studies, Conflict Resolution, and Peace Policies, and another Master's degree in Development Communication. He has also obtained several certifications in international human rights law and transnational organized crime. He is currently a doctoral candidate, studying the impact of new digital technologies on international security, including cyber operations and hybrid warfare. He is affiliated with the Centre for African Studies at Leiden University. Previously, he worked for five years on the implementation of a reintegration programme for former Boko Haram child soldiers in Niger.

18 UN Security Council, Resolution 1540 (2004), S/RES/1540.


CYBER ATTA
CYBER ATTA
NEWS NEWS NEWS
CYBER Att

The P3 equation: cyber warfare shaping peace, power & perception

by Sarra Hannachi

Cybersecurity: the human in the cyber

Human security as a concept has always been disputed, largely because every new "security" context in history has contributed to its meaning. Traditionally, human security has been understood in terms of physical violence and instability, associated with being protected from armed conflict, crime, terrorism, and extremism, while framing security within the boundaries of military-state paradigms.

Today, that concept has evolved because we live in a digitally transformative era defined by rapid technological development that is accessible both intellectually and financially. Like any resource in the world, its utility is determined by how humans deploy it. This is how the duality of technology manifests in an era of rising geopolitical tensions, crimes, and violations. As much as digital technologies in cyberspace can serve infrastructure and the public sector, they can also be weaponized – like chemicals or steel. There is no human security without cybersecurity.[1]

Peace under digital siege

Peacebuilding and peacekeeping efforts for human security continue to be threatened by

1 Fukuda-Parr, S., & Messineo, C. (2012, January). Human Security: A Critical Review of the Literature (Working Paper No. 11). Centre for Research on Peace and Development (CRPD). https://sakikofukudaparr.net/wp-content/uploads/2013/01/HumanSecurityCriticalReview2012.pdf

a digital siege composed of cyberattacks, surveillance spyware, data theft and manipulation, digitally generated or disseminated disinformation campaigns, and other misleading operations. Peace efforts have been undermined by such threats, including, for example, intentional digital infrastructure sabotage – such as the attacks in Ukraine that disrupted telecommunications and Internet connectivity[2] – as well as recent ransomware incidents in Iran, which stole and gradually leaked sensitive public-sector and banking data.[3]

These cyberattack operations play a role in cyberwarfare and cyber terrorism and can jeopardize peace agreements during war, as well as due-diligence processes in digital security.

Cyber warfare creates a self-reinforcing cycle capable of disrupting peace without leaving physical traces or legally binding evidence in some cases, making accountability for cyber operations during war significantly more complex and dividing stakeholder perspectives

When digital forensics reveal that what once appeared to be fact is a manipulated narrative, it becomes significantly more difficult to undertake diplomatic tasks such as mediation or drafting terms for accountability, peace, or ceasefire agreements. It is also difficult when human security assessments cannot effectively take place because cyberattacks damage infrastructure, mobility, resources, and vital sectors.

Power dynamics in cyber weaponization

State actors are no longer the only entities using their technological advancement for dig-

2 Frąckiewicz, M. (2025, June 21). Telecommunications Infrastructure in Ukraine (2022–2025): Destruction and Resilience. TS2 Space. https://ts2.tech/en/telecommunications-infrastructure-in-ukraine-2022-2025-destruction-and-resilience/

3 Olney, M. (2025, June 30). 5 of the biggest cyber-attacks of 2025 (So far). Industry Trends & Insights. https://insights.integrity360.com/5-of-the-biggest-cyber-attacks-of-2025-so-far

ital deterrence and non-physical control of the battlefield alongside traditional military forces. Non-state actors involved in cyber operations during war or conflict can also obstruct the international community's pursuit of peace and justice by deliberately threatening the human security of both combatants and civilians aligned with the opposing party.[4]

It shows how fragile global peace and security are in digital spaces, and unpredictable diplomatic outcomes can be in this environment. Furthermore, cyber operations may not be entirely dependent on technical experts due to the rapid availability of open-access intelligence and technologies, but they remain dependent on digital tools.

Weaponizing these tools and transforming cyberspace into a battlefield have significantly influenced the outcomes of warfare and become highly contingent on the digital economy.[5] The political size of actors matters less as cyber tactics alone can redefine leverage.

> **State actors are no longer the only entities using their technological advancement for digital deterrence and non-physical control of the battlefield alongside traditional military forces**

@UN Photo/ Evan Schneider

4 Borghard, E. D., & Lonergan, S. W. (2019). Cyber Operations as Imperfect Tools of Escalation. Strategic Studies Quarterly, 13(3), 122–145. Available at: https://www.jstor.org/stable/26760131

5 Rojas, V., & Lissette, A. (2023, September 22). The Rise of Net-States in the Cyberspace: Cyber Power Dynamics and the Disruption of International Security. Digitální Repozitář UK. https://dspace.cuni.cz/handle/20.500.11956/187380


Credit: UNFICYP

Perception as a trust-building factor

Historically, there are methods for building trust and confidence between states and their military units, which include Confidence and Security Building Measures (CSBMs). They are traditional tools used to foster trust between states and their military units. The Organization for Security and Co-operation in Europe (OSCE) leads confidence-building efforts among its participating states,[6] while other international organizations such as the United Nations Office for Disarmament Affairs (UNODA) also implement CBMs in the broader context of arms control and global security [7] to ensure new developments in capacities – such as biological or technological innovation – are not additionally exploited for the modernization and proliferation of arms.

While CBMs have proven effective for some states by allowing yearly field visits, regular military information exchanges, and other commitments, transparency in information exchange cannot be guaranteed because visits remain limited to physical military spaces, resources, and equipment, disregarding the intangible developments that may result in cyberwarfare. This leads to trust built on a perception rather than on full compliance with international frameworks.

Pathways to cyber-enabled justice & security

Cyber warfare creates a self-reinforcing cycle capable of disrupting peace without leaving physical traces or legally binding evidence in some cases, making accountability for cyber operations during war significantly more complex and dividing stakeholder perspectives.[8]

Integrated approaches are therefore key to global stability, cybercrime mitigation and human security. Combining technology, diplomacy, and security can provide all the information and resources needed to mediate peace, de-escalate tensions, seek justice for actions that machines cannot be held responsible for, and balance geopolitical and

6 History and Background of Confidence- and Security-Building Measures (CSBMs) in the OSCE. (2004). OSCE. https://www.osce.org/fsc/40035

7 Military confidence building measures | United Nations Office for Disarmament Affairs. https://disarmament.unoda.org/en/our-work/cross-cutting-issues/military-confidence-building-measures

8 Roudani, C. (2025, June 18). Cyber Deterrence and Digital Resilience: Towards a New Doctrine of Global Defense. Modern Diplomacy. https://moderndiplomacy.eu/2025/06/18/cyber-deterrence-and-digital-resilience-towards-a-new-doctrine-of-global-defense/

cyber-power dynamics.[9] This is achieved not only through multi-stakeholder strategies in peacebuilding processes, ethical technology deployment during conflict, and cyber defense mechanisms, but also through far-sighted planning. Hence, it is important to invest in proactive CBMs in cyberspace and digital platforms as much as in physical weaponry and systems, to prevent conflict escalation and the spread of disinformation.

Overall, embedding cyber resilience into broader peace and security frameworks to ensure justice, stability, and trust forms the core of what this article refers to as the P3 equation.

> **State actors are no longer the only entities using their technological advancement for digital deterrence and non-physical control of the battlefield alongside traditional military forces**

ABOUT THE AUTHOR

Sarra Hannachi is a Mozilla Fellow and a member of the Coalition for Independent Technology Research. She collaborates with non-governmental organizations and think tanks to advance peace and security through the ethical development, use and policymaking of AI and emerging technologies. Sarra specializes in evidence-based research and technology policy, particularly in border security and contexts of war and armed conflict, drawing on experience in the Middle East and North Africa (MENA) region and Africa to provide strategic recommendations on responsible AI in security settings. For youth representation and empowerment, Sarra also serves as a member of the African Union-European Union (AU-EU) Youth Lab Advisory Board with the AU-EU Youth Voices Lab initiative.

9 Van der Meer, S. (2015). Enhancing International Cyber Security: A Key Role for Diplomacy. Security and Human Rights, 26(2–4), 193–205. https://doi.org/10.1163/18750230-02602004


The line between the digital and the physical world is being crossed with increasing frequency
CITY1 //// MAINCITY
750 // +1000
600 // +0

Cyberterrorism: legal and governance challenges

by Yéelen Marie Geairon

As the digital landscape evolves at unrelenting speed and global interconnectivity deepens, terrorist groups are increasingly exploiting these dynamics with novel approaches, making cyberterrorism one of the most significant and complex threats of our time. Unlike traditional terrorist threats, bound by physical borders and material means, cyberterrorism transcends these limits, eroding security capacities, straining legal systems, and testing the resilience of international governance.

Cyberterrorism can be defined as the unlawful use of information and communication technologies (ICTs) to intimidate or coerce populations or governments for political, ideological, or religious purposes. It may take the form of cyber-attacks directly targeting computer systems, or the use of ICTs to facilitate the commission of terrorist offences, such as propaganda, financing, planning, or carrying out attacks.[1] Hybrid by nature, cyberterrorism often mixes traditional forms of terrorism, such as physical attacks or acts of sabotage, with digital forms of attack, such as hacking systems to gather sensitive information, executing ransomware attacks or disrupting digital or physical infrastructure. The line between the digital and the physical world is being crossed with increasing frequency, giving terrorists alternative

1 United Nations Office on Drugs and Crime, "Cyberterrorism", Education for Justice (E4J) University Module Series on Cybercrime, Module 14, Key Issues. https://www.unodc.org/e4j/zh/cybercrime/module-14/key-issues/cyberterrorism.html

methods to execute some or all parts of their crimes, as well as disrupting investigative or preventative efforts.

In this context, tools such as generative artificial intelligence (AI) can facilitate the production of falsified content, such as *deepfakes*, to incite violence or recruit, while the *dark web* can serve as a platform for exchanging sensitive data, ranging from anonymized networking and planning of attacks, to classified information obtained through hacking and procurement of weapons (e.g., plans for 3D-printed weapons). Tools initially designed to protect privacy and the security of communications are thus being misused by terrorist groups to conceal their activities and fund their operations. AI further enables the automation of *phishing* campaigns, the tailoring of propaganda to specific target groups, and the exploitation of public information to identify technical vulnerabilities or potential targets.

In light of the rapid evolution of these criminal capabilities, recent analyses, notably by the United Nations[2] and the Global Centre on Cooperative Security,[3] have warned against terrorist groups using emerging technologies to target state infrastructure, particularly military, diplomatic, energy, and civil institutions. This threat has been identified by Europol in the European Union Terrorism Situation and Trend Report 2025,[4] in which the rise of terrorist cyber threats targeting critical infrastructure represents a growing concern.

The fight against cyberterrorism faces a fragmented normative landscape, marked by the absence of a universal definition of terrorism and cyberterrorism

These observations reinforce the fact that this threat cannot be measured solely in terms of its immediate impacts, but also in terms of how it challenges current legal frameworks and mechanisms governing international cooperation.

In fact, the fight against cyberterrorism faces a fragmented normative landscape, marked by the absence of a universal definition of terrorism and cyberterrorism, as well as gaps in the rules applicable to non-state actors in cyberspace. This results in varying legal thresholds that hinder extradition, mutual legal assistance, and cross-border evidence sharing, particularly when national laws differ on the definition of offences or on double criminality requirements.

Moreover, although the Budapest Convention on Cybercrime remains a reference instrument for criminal cooperation and the preservation of electronic evidence, it is not binding on States that are not party to it, thus limiting its effectiveness on a universal scale. Similarly, United Nations Security Council Resolutions 1373 (2001) and 2462 (2019) require States to criminalize terrorist acts and disrupt their financing, including when such acts involve ICTs or virtual assets. This dimension is particularly relevant in the context of cyberterrorism, where digital platforms and cryptocurrencies can facilitate the collection and transfer of funds. Yet again, the lack of legal harmonization and the differing technical capacities of States continue to hinder the uniform application of these obligations, weakening the collective response.

2 United Nations Office of Counter-Terrorism, 2023 Counter-Terrorism Week Report, April 2023. https://www.un.org/counterterrorism/en/2023-counter-terrorism-week

3 Global Center on Cooperative Security, Blue Sky VI: An Independent Analysis of UN Counterterrorism Efforts, June 2023. https://www.globalcenter.org/wp-content/uploads/Global-Center_Blue-Sky-VI-Report_June-2023.pdf

4 https://www.europol.europa.eu/cms/sites/default/files/documents/EU_TE-SAT_2025.pdf

While significant, these initiatives fail to address the persistent gaps in the legal framework applicable to non-state actors in cyberspace, particularly terrorist groups. This has fuelled the work of the Ad Hoc Intergovernmental Committee under the United Nations Office on Drugs and Crime, which led to the adoption of the International Convention against the Criminal Use of Information Technologies by the General Assembly in 2024.

Pending its entry into force and in the face of the pressing threat, the response to cyberterrorism therefore still relies largely on national measures, which are often designed according to each state's own security priorities.

The response to cyberterrorism therefore still relies largely on national measures

While they can allow for a rapid response, these isolated approaches present two major risks. First, they undermine the international cooperation necessary for effective attribution of attacks; second, they may, in the absence of safeguards, disproportionately restrict fundamental rights such as privacy or freedom of expression. This dual vulnerability underlines the importance of articulating operational efficiency and compliance with international standards, in particular through independent and harmonized control mechanisms.

Furthermore, ensuring that measures comply with fundamental rights is only one part of the response. Another key dimension is the effective capacity of States to investigate and prosecute the perpetrators of cyberterrorist attacks. Assigning responsibilities remains one of the major challenges, especially due to the use of relay servers, anonymization technologies and infrastructures located in third-party jurisdictions. To address this issue, national capacities must be supported by solid technical expertise, advanced tools, as well as common protocols to ensure the admissibility of evidence collected before the courts. In recognition of these challenges, Member States stressed, in Security Council resolution 2341 (2017) and in the United Nations Global Counter-Terrorism Strategy, the importance of multi-stakeholder cooperation, involving international, regional, and subregional organizations, the private sector and civil society.[5]

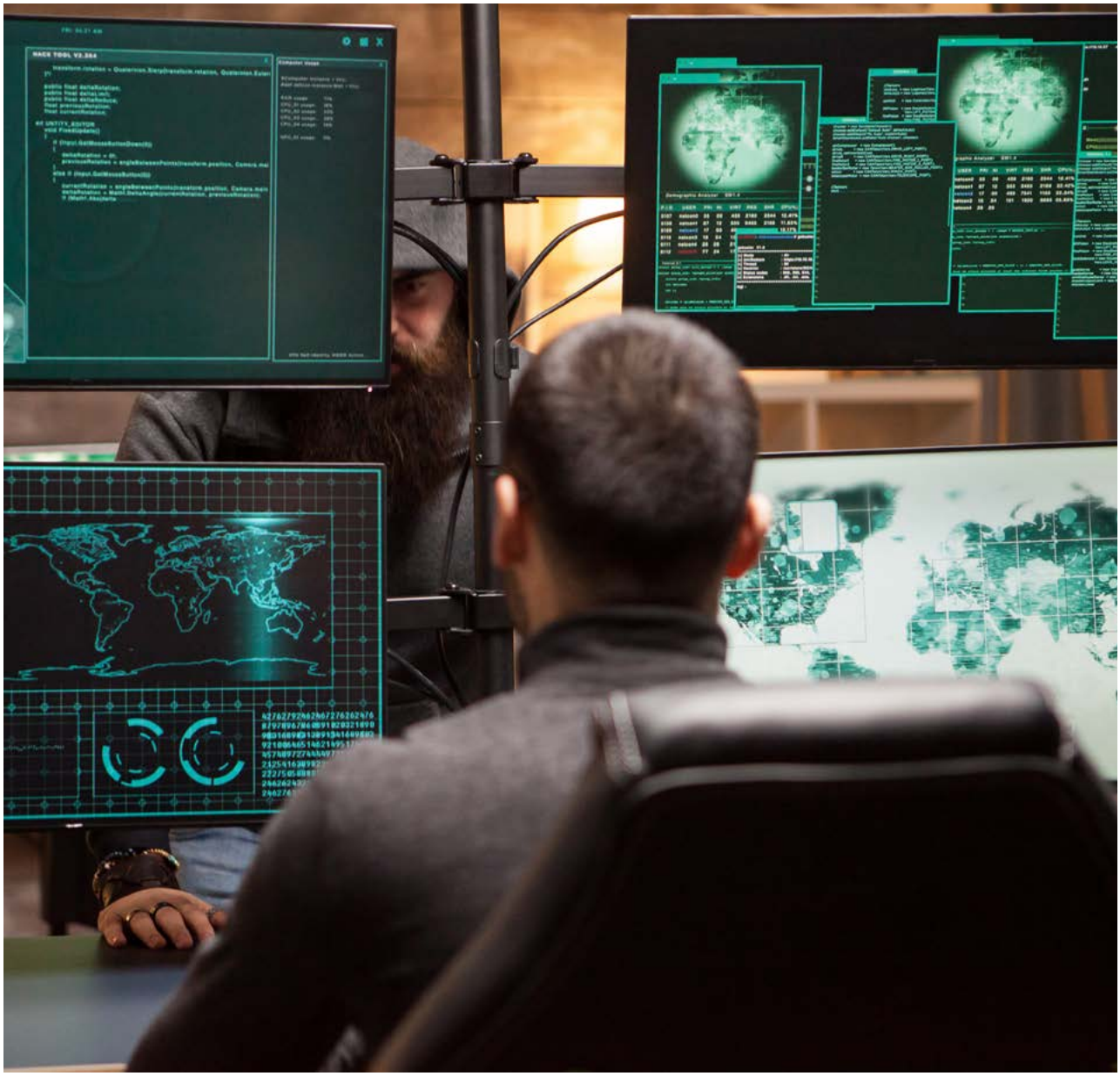
Strengthening technical capacities and international cooperation is not only aimed at identifying and punishing perpetrators, but also at ensuring that victims receive effective protection. Indeed, the human rights dimension of the response to this threat implies considering access to justice and reparation for victims of cyberterrorist attacks, whether they are direct victims of disruptions to vital services or people affected by breaches of their personal data.

Procedures must enable them to assert their rights and obtain redress, which requires legislative and institutional frameworks adapted to the transnational and technical nature of attacks.

Hence, an effective response to cyberterrorism requires the development of integrated legal frameworks linking anti-terrorism laws and cybercrime legislation, expanded international cooperation, proactive governance of emerging technologies, and a substantial strengthening of investigative and prosecutorial capabilities.

5 During the 8th review of the Global Counter-Terrorism Strategy, the General Assembly requested "the Office of Counter-Terrorism and other relevant Global Counter-Terrorism Coordination Compact entities to jointly support innovative measures and approaches to build the capacity of Member States, upon their request, for the challenges and opportunities that new technologies provide, including the human rights aspects, in preventing and countering terrorism".

In the face of this rapidly evolving threat in cyberspace, agility, cooperation, and the rule of law remain essential. Accordingly, meeting the challenge of cyber-terrorism requires moving beyond fragmented national approaches to build a truly global approach that can adequately balance state sovereignty, security imperatives, and respect for fundamental rights.



ABOUT THE AUTHOR

Yéelen Marie Geairon is a legal expert in counter-terrorism, defence policy, and international criminal law, with experience in governmental, parliamentary, and international organizations, including the United Nations Office of Counter-Terrorism in Rabat (Morocco), Eurojust, and the Council of Europe. Her work focuses on the intersection of security, justice, and human rights, with particular expertise in the governance of terrorism threats and associated security challenges, as well as emerging risks.

unicri
United Nations Interregional Crime and Justice Research Institute

LUMSA HUMAN ACADEMY

SUMMER SCHOOL

ON ARTIFICIAL INTELLIGENCE (AI), ETHICS AND HUMAN RIGHTS

Join

the **United Nations Interregional Crime and Justice Research Institute** and **LUMSA Human Academy** for a dynamic hybrid program to explore the ethical challenges at the intersection of AI and human rights.

Gain understanding of the impact of AI on society, public safety, personal freedom and marginalized communities.

Info

- **Expert insights.**
- Practical **exercises** and **simulations.**
- Real-world **case studies.**
- Certificate of Participation issued by **UNICRI** and **LHA.**

This hybrid program allows for all students to have an immersive and adaptable experience, in a truly global classroom.

English

Hybrid (Rome or Online)

Students, Post-Graduates and Professionals

Foster a human-centric approach to AI, shape the future of technology and human rights!

unicri
United Nations Interregional Crime and Justice Research Institute

10th edition

SUMMER SCHOOL

ON MIGRATION AND HUMAN RIGHTS

Join

the **United Nations Interregional Crime and Justice Research Institute** and **John Cabot University** in Rome, Italy, for an immersive experience focusing on the intersection of migration, security and human rights.

Gain the legal, geopolitical and institutional tools needed to navigate one of today's most urgent global challenges.

Info

- **Expert insights.**
- Practical **exercises** and **simulations.**
- Real-world case studies.
- Certificate of Participation issued by **UNICRI** and **JCU.**

The in-person format is ideal for those seeking to deepen their knowledge and network in the field of migration.

English

In-Person at JCU (Rome)

Students, Post-Graduates and Professionals

Join the conversation. Shape global migration policy with respect for human rights!

IN FOCUS

by Alliou Traoré

From persecution to peace: the resilience of Hamawi Sufis as a model for preventing violent extremism in the Sahel

In a Sahelian context marked by the rise of violent extremism, the Hamawi Sufi community of Yacouba Sylla in Kaédi (Mauritania) offers a remarkable example of collective resilience.

This Sufi brotherhood has preserved its cohesion, identity, and practices over time, despite considerable external pressures: from colonial persecution to contemporary challenges linked to extremism (Cissé, 2005; Hanretta, 2003; Traoré, 2019). How have the members of this community managed to maintain their bonds and pass them on to new generations? This article highlights key historical, social, and spiritual mechanisms that have enabled this Sufi brotherhood to endure.

The Hamawi Sufi community of Yacouba Sylla in Kaédi

Yacouba Sylla is described as a devoted disciple of Chérif Hamahoullah. His bond with the spiritual master was based on deep respect and full commitment to the principles of the Tijaniyya Hamawiyya order.

However, while Sylla greatly honoured Cheikh Hamahoullah, he preferred to engage directly with his spiritual legacy rather than through intermediaries or hierarchical structures.

He emerged as a charismatic figure capable of mobilising followers around a clear vision. His spiritual authority stemmed from his deep knowledge of Hamahoullah's teachings and his ability to adapt them to diverse contexts.

The Hamawi Sufi movement led by Yacouba Sylla began in 1929, culminating in the events of February 1930. At the time, according to estimates by Sean (2023), the number of followers of this religious movement exceeded 600, "frequently in conflict, and increasingly violently, with other residents of Kaédi."

These events led to a wave of violence against the leaders and followers of this religious movement. Yacouba Sylla was placed under house arrest in 1930 in Sassandra for eight years; Chérif Hamahoullah was transferred from his internment in Mederdra, Mauritania, to Adzopé in 1935; and many followers were arrested, deported, imprisoned, or even killed. Despite a wounded history and longstanding conflict with the colonial authorities, the Hamawi Sufi community of Yacouba Sylla continued to strengthen itself, drawing on this past not as a reason for revenge but as a catalyst of resilience and spiritual redemption. In a region of the Sahel plagued by extremist violence, this community has stood out for its non-violence.

Today, the community faces new challenges: some youth become distanced from the community, external ideological influences, evolving traditional lifestyles, and sometimes a lack of economic opportunities (Ebode & Njoya, 2023; RAND Corporation, 2020). Nevertheless, it persists. Youth involvement in spiritual activities remains remarkable. Even those who temporarily move away for studies or work remain deeply connected to the community and its values. This persistent connection is a significant protective factor against radicalisation (Cachalia et al., 2016; Center on Global Counterterrorism Cooperation, n.d.).

Resilience: a multidimensional concept

Resilience, far from being limited to individual psychology, is now understood as a dynamic process involving personal, social, and institutional factors. According to Herrman et al. (2011), it represents "positive adaptation to adversity, an interactive process influenced by personal, biological, and social factors." Grossman (2021) expands this to a "multi-systemic process involving psychological, educational, and community systems."

This holistic approach is particularly relevant for understanding the community resilience of Hamawi Sufis
As Folke (2016) points out, resilience is "the capacity of systems to persist, adapt, or transform in response to dynamic and

unexpected changes." Adger (2000) notes that it "depends on institutions, social networks, and trust-based relationships that enable individuals to cooperate in overcoming disruptions."

Three factors of resilience stand out in the case of the Hamawi Sufis:

1. A foundation of core values as an identity anchor

Interviews with community leaders and youth reveal that their resilience rests on three core values: fear of God (taqwa), adherence to religious precepts, and avoidance of prohibited behaviours. These principles shape not only individual conduct but also collective cohesion (Cissé, 2000; Hamès, 1983).

"These values are not just abstract concepts," explains one community imam. "They form a holistic moral framework that governs our daily interactions and shields us from harmful external influences." A young leader adds: "When these values are deeply internalised, one naturally develops resistance to extremist narratives that advocate violence."

Collective spiritual practices reinforce these values. Hadrat (group invocations) and qasidas (religious chants) create

"Youth involvement in spiritual activities remains remarkable. Even those who temporarily move away for studies or work remain deeply connected to the community and its values

intense moments of communion, strengthening a sense of belonging. These daily rituals are far from mere traditions; they act as effective barriers to the social isolation often exploited by extremist recruiters.

2. Collective memory and mystical reinterpretation of trials

One particularly striking aspect of Hamawi resilience lies in their relationship with history. Stories of past persecution: Cheikh Hamahoullah's imprisonment, the February 1930 shootings in Kaédi, the November 1941 events in Yelimane, the forced exile of Yacouba Sylla, are passed down to younger generations not as causes for resentment but as testimonies of spiritual fidelity (Hanretta, 2023; Traoré, 2019).

This mystical reinterpretation of adversity transforms collective traumas into sources of spiritual strength rather than justifications for radicalization (Schmitz, 1985; Robinson, 2000). Intergenerational education and the preservation of strong collective memory thus serve as essential protective factors. The practice of endogamy, though less strict today, also historically helped maintain community cohesion and safeguard core values (Koita, 2016).

3. Intergenerational transmission mechanisms

The transmission of values occurs through a blend of ancestral traditions and modern means (such as audio/video recordings, and social media). Families play a central role, complemented by community gatherings and the influence of spiritual leaders such as the khalifes (Jourde, 2009; Triaud, 2010). Audio and video recordings of sermons and hadrat allow young people to access teachings even when physically distant. Generational structures (known as fedde in Soninké, informal youth associations fostering peer mentoring) ensure horizontal peer transmission that effectively complements the vertical transfer from elders to youth.

The experience of the Hamawi Sufi community of Yacouba Sylla reminds us that community resilience relies on a living synergy of shared values, safeguarding practices, and adaptive responses to context

This transmission is even more crucial today, amid strong ideological influences. Some community members have expressed concern about youth who are more vulnerable to outside influences. In response, the community has developed specific initiatives: educational talks tailored to youth concerns, strengthened intergenerational ties, and the creation of a talent database to support professional integration. Cultural and sporting events, along with intergenerational discussions, are also encouraged to foster solidarity and identity anchoring.

Recommendations for preventing extremism

Lessons drawn from the Hamawi experience can be applied to other Sahelian and African contexts. These include promoting shared values, cultivating an inclusive collective memory, strengthening intergenerational bonds, and mobilizing youth as resilience actors (UNOCT, 2022; USAID, 2024). Integrating these principles into public policy and education programmes can help prevent radicalization and promote peace.

This study highlights several concrete initiatives:

- Establishing intergenerational dialogue spaces, where young

people and elders share life stories and spiritual teachings.

- Organising sessions on community history to build strong cultural and spiritual identity.
- Emphasising the education of women and girls, traditionally less involved in hadrat, to ensure value transmission within families.
- Expanding educational talks and cultural and sporting events to foster youth engagement and community cohesion.

To further strengthen resilience, it is recommended to:

- Value local cultural and spiritual resources in prevention programmes.
- Strengthen intergenerational transmission through modern tools and dialogue spaces.
- Engage youth as resilience actors by supporting their leadership and participation.
- Integrate values education into formal curricula by developing appropriate teaching modules.

- Use collective memory as a resource by encouraging constructive engagement with history.

Conclusion

The experience of the Hamawi Sufi community of Yacouba Sylla reminds us that community resilience relies on a living synergy of shared values, safeguarding practices, and adaptive responses to context. This people-centred, community-based model offers practical lessons for designing effective interventions in communities vulnerable to extremism, thus fostering peace, stability, and sustainable development. The approaches emerging from this study provide concrete solutions for implementing resilience programmes to counter violent extremism across various environments. By promoting identity anchoring and solidarity dynamics, such programmes can prevent individual vulnerability to radical influences.

> **When these values are deeply internalised,one naturally develops resistance to extremist narratives that advocate violence**

Key References

AUC, UNDP & UNRISD (2024). Roots of African Resilience: *A Transformative Approach*.

Cissé, C. (2000, 2005). *Cheikh Yacouba Sylla et le Hamallisme en Côte d'Ivoire ; La confrérie hamalliste face à l'administration coloniale française*.

Ebode, J. V. N., & Njoya, H. N. (2023). *Défis et bilan de la lutte contre le djihadisme en Afrique*.

Folke, C. (2016). *Resilience (Republished)*.

Grossman, M. (2021). *Resilience to Violent Extremism and Terrorism: A Multisystemic Analysis*.

Hamès, C. (1983). *Cheikh Hamallah ou Qu'est-ce qu'une confrérie islamique (Tarîqa)?*

Hanretta, S. (2003, 2023). *Constructing a Religious Community in French West Africa ; Genre et capacité d'action dans l'histoire d'une communauté soufie d'Afrique de l'Ouest*.

Herrman, H. et al. (2011). *What is Resilience?*

Holling, C.S. (1973). *Resilience and Stability of Ecological Systems*.

Koita, T. (2016). *Kaédi, la ville éternelle*.

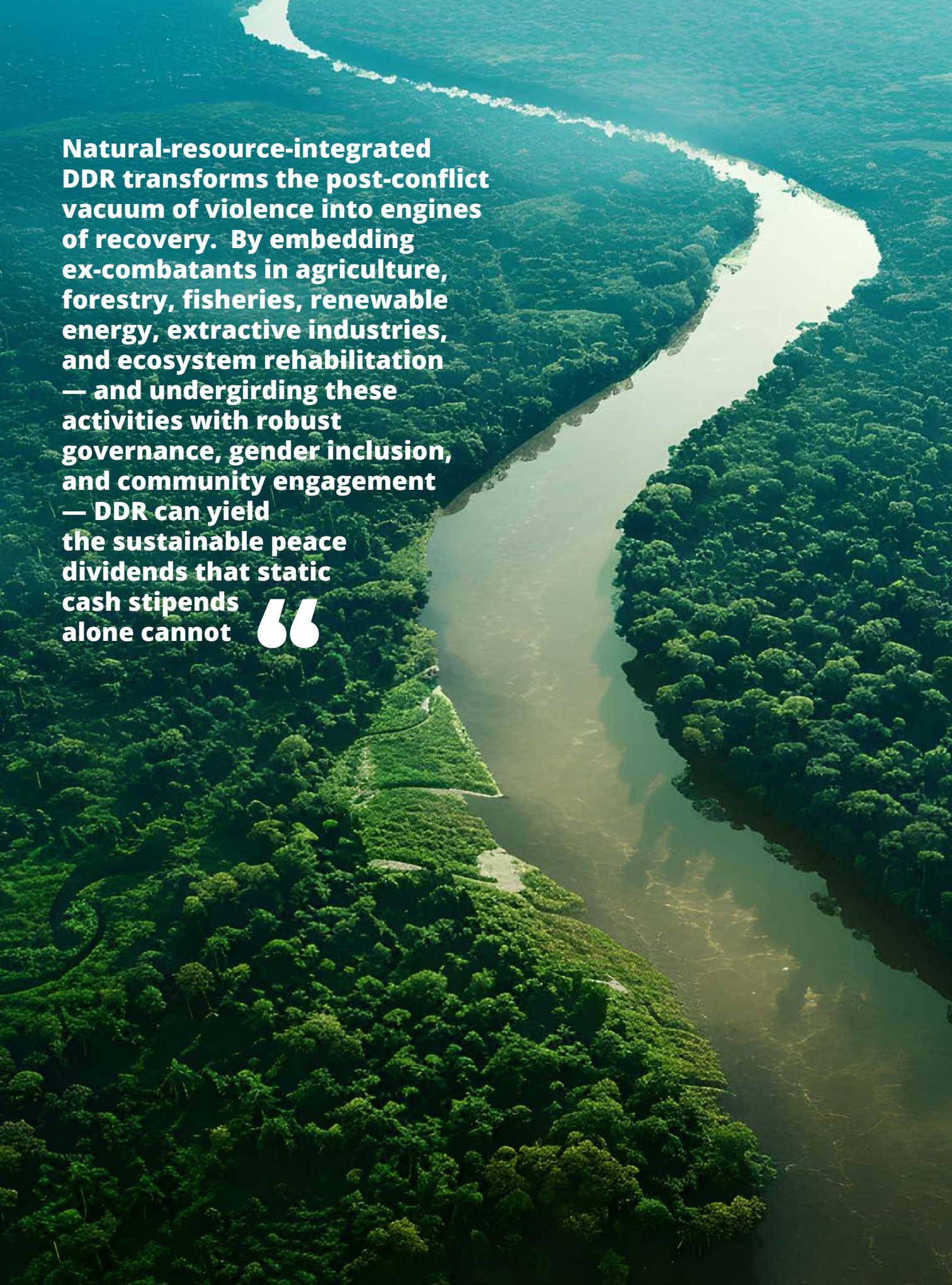
RAND Corporation (2020). *Countering violent extremism: A review of the evidence*.

Robinson, D. (2000). *Paths of accommodation: Muslim societies and French colonial authorities in Senegal and Mauritania, 1880–1920*.

Traoré, A. (2019). *Cheikh Hamahoullah: homme de foi et résistant ; l'islam face à la colonisation française en Afrique de l'Ouest. L'Harmattan*.

ABOUT THE AUTHOR

Alliou Traoré is an international development expert and peacebuilding practitioner with over 15 years of experience across West Africa. He specializes in community resilience, conflict transformation, and preventing violent extremism. His work combines field-based research with a deep engagement in spiritual and cultural dynamics in fragile contexts. He is the author of several studies on Sahelian Islam and resilience narrative.



Natural-resource-integrated DDR transforms the post-conflict vacuum of violence into engines of recovery. By embedding ex-combatants in agriculture, forestry, fisheries, renewable energy, extractive industries, and ecosystem rehabilitation — and undergirding these activities with robust governance, gender inclusion, and community engagement — DDR can yield the sustainable peace dividends that static cash stipends alone cannot

IN FOCUS

by Cristian Mazzei

Reintegrating Peace Through Natural Resources: Enhancing DDR for Sustainable Stability

Post-conflict societies often hinge on the success of Disarmament, Demobilization, and Reintegration (DDR) programmes to transition former combatants into productive civilians. Yet, DDR too frequently stalls when it comes to securing long-term livelihoods. Ex-fighters, lacking viable economic alternatives, may cling to arms to protect illicit resource exploitation — perpetuating cycles of violence. My research highlights that embedding natural-resource-based livelihoods into DDR not only aligns income generation with environmental stewardship but also establishes the durable foundation that peace operations desperately need.

The reintegration of ex-combatants is not a quick fix but a long-term process that unfolds at individual, communal, and national levels. Programmes must avoid the perception of "rewarding" fighters with handouts; instead, they should foster lasting income, social belonging,

and political participation.[1] In eastern DRC's UNDP Community Recovery and Reintegration Programme, 39 percent of North Kivu participants depended directly on natural resources for their livelihood, compared with just 20 percent in South Kivu.[2] This disparity underscores how ineffective resource governance can incentivize combatants to remain armed — using weapons to secure illicit mining, logging, or water-control rackets.

However, DDR faces two principal obstacles that must be addressed explicitly. First, the lack of equitable control over land, water, and resource markets often leaves ex-combatants marginalized, reinforcing distrust and stalling reintegration. Second, when economic and political opportunities remain concentrated in the hands of the victorious faction, many former fighters view DDR as hollow, prompting a return to arms in search of survival.

To counteract these challenges, DDR must integrate natural-resource activities that ex-combatants can legally exploit. Agriculture and animal husbandry, for example, are prominent reintegration pathways across Afghanistan, Angola, Liberia, and Rwanda. In Afghanistan's New Beginnings Programme, former soldiers received seed kits and tools but lacked land-tenure support or modern training; still, 43 percent chose farming within

1 Mazzei, C., Ware, H. & Vom Strokirch, K., *Environmental Peacebuilding and a UN Rapid Response Force as a Counter to Conflicts over Natural Resources* (PhD thesis, 2025), ch. 6.

2 United Nations Development Programme, Community Recovery and Reintegration Programme: North Kivu, Democratic Republic of the Congo (Final Evaluation Report, UNDP, 2018).

six months.[3] In Rwanda's first DDR phase, 30 percent of participants secured formal employment, 40 percent became self-employed in agriculture, and the remainder engaged in small-scale farm work.[4] These figures demonstrate both the potential and the pitfalls of cash-and-tools approaches when they lack robust property rights and market access.

Forests and non-timber forest products (NTFPs) present another avenue. Well-managed timber cooperatives and Payments for Ecosystem Services can provide ex-combatants with income while preserving critical ecosystems.[5] Women — who shoulder the daily burden of collecting fuelwood and wild foods — benefit particularly from gender-sensitive forestry initiatives that diversify livelihoods and distribute forest-management roles more equitably. Likewise, water and sanitation projects — such as Darfur's community wells and Sierra Leone's rehabilitation of irrigation schemes — have proven to unite rival groups, improve public health, and generate employment for ex-fighters.[6]

> **"The lack of equitable control over land, water, and resource markets often leaves ex-combatants marginalized, reinforcing distrust and stalling reintegration**

Renewable energy solutions further extend DDR's reach into natural-resource sectors. Off-grid solar lantern programmes, improved cook-stove manufacturing and micro-hydro installations offer technical training for ex-combatants, reduce reliance on biomass fuels, and light schools and clinics in remote areas.[7] In the mining sector, linking DDR to artisanal cooperatives helps stem the tide of criminal exploitation. Despite the DRC's estimated US $24 trillion in mineral wealth, a $2 billion copper investment generated fewer than 3 000 formal jobs — highlighting the urgency of designing resource-industry partnerships that deliver real employment for ex-combatants.[8]

Coastal and aquatic resources also play a vital role. In Sierra Leone and Liberia, community fishing cooperatives and regulated landing sites boosted ex-combatant incomes by around 30 percent while sharply reducing re-recruitment.[9] Aceh's small-scale aquaculture ponds for milkfish and tilapia offered inclusive work for both men and women, demonstrating how ecosystem-based livelihoods can reinforce social cohesion.[10] Protected areas likewise serve as platforms for reintegration: Mozambique's Gorongosa National Park rehired 74 former combatants as park rangers in 1994, combining biodiversity recovery with the restoration of law and order.[11]

Perhaps the most emblematic example comes from Afghani-

3 United Nations Development Programme, *Afghanistan New Beginnings Programme: Demobilisation and Reintegration of Ex-Soldiers* (UNDP, 2004).
4 Rwanda Demobilization and Reintegration Commission, *National DDR Programme: First Phase Report* (Kigali, 2001).
5 UN Environment Programme & UNDP, *The Role of Natural Resources in Disarmament, Demobilization and Reintegration* (UNEP/UNDP, 2009).
6 Mazzei et al., ibid., "DDR in Water and Sanitation."
7 Mazzei et al., ibid., "DDR in EnergyX Sector."
8 Mazzei et al., ibid., "DDR in Mining and other Extractive Industries."
9 Mazzei et al., ibid., "DDR in Fisheries (Wild and Aquaculture)."
10 Ibid.
11 Mazzei et al., ibid., "DDR in Protected Areas and Ecotourism."

stan's Afghan Conservation Corps (ACC). Established in 2003, the ACC engaged ex-combatants in reforestation of pistachio woodlands and conifer forests, restoring 108 nurseries and planting 150,000 conifers and 350 000 fruit trees annually. By 2009, these efforts had generated 400,000 labour-days, simultaneously rebuilding ecosystems and forging meaningful employment.[12] Such large-scale ecosystem restoration projects deliver a triple dividend: income, environmental resilience, and community reconciliation.

Yet even when DDR programmes succeed initially, civil conflicts can reignite if resource benefits remain unevenly distributed. When a single group controls mining revenues, forest concessions, or water rights, it consolidates power and sows the seeds of new grievances — undermining reconciliation and setting the stage for renewed violence.

When a single group controls mining revenues, forest concessions, or water rights, it consolidates power and sows the seeds of new grievances — undermining reconciliation and setting the stage for renewed violence

Successful reintegration hinges on community support and shared ownership. DDR programmes that include ex-combatants in natural-resource management councils and community dialogues affirm their civilian status and foster reconciliation between former fighters, local populations, and the state. This participatory approach is especially crucial for women: in Côte d'Ivoire, women comprised 8 percent of the 74,000 DDR beneficiaries, while in Colombia's 2016 Peace Agreement, one-third of the more than 13,000 ex-combatants were women.[13]

Ultimately, natural-resource-integrated DDR transforms the post-conflict vacuum of violence into engines of recovery. By embedding ex-combatants in agriculture, forestry, fisheries, renewable energy, extractive industries, and ecosystem rehabilitation — and undergirding these activities with robust governance, gender inclusion, and acommunity engagement — DDR can yield the sustainable peace dividends that static cash stipends alone cannot.

ABOUT THE AUTHOR

With two decades of UN experience, **Cristian Mazzei** is a seasoned professional in international peace and security, human rights, and governance. His roles have spanned political affairs, policy advising, planning, and program management at headquarters and in the field. As the author of the widely praised Special Assistant's Handbook, Mazzei provides practical insights for UN Special Assistants, enhancing their effectiveness and aiding senior management in optimizing office operations. He has made significant contributions to peace and security in diverse countries, including Haiti, the Democratic Republic of Congo, Lebanon, the Central African Republic, and Somalia. His work with UNICEF and UNEP, along with insights gained from UN Reform efforts in the Central African Republic, reflect his commitment to global development. Direct engagement with local communities has been a cornerstone of Cristian's career, providing him with a deep understanding of their realities and motivating his efforts to alleviate suffering.

12 Afghan Conservation Corps, *Annual Report 2009* (Ministry of Agriculture, Afghanistan, 2010).

13 Mazzei et al., ibid., "Disarmament, Demobilization and Reintegration."

The human element remains a significant challenge to cybersecurity resilience, and despite technological advancements, this area is still heavily neglected

Examining the hidden risk in cyber conflict: human behaviour as the critical blind spot

by Christopher Weir and Ally Zlatar

Introduction

Despite significant advancements in cybersecurity technology and defensive strategies, human errors continue to be a major vulnerability. This is often due to negligence or a lack of education and awareness, making the human element a persistent blind spot for organizations around the world. This paper examines the role of human behaviour in cybersecurity breaches by utilising case-study analysis to highlight how several key human-caused cyber-attacks, such as social engineering, poor security measures, and insider threats, continue to undermine even the most heightened security infrastructures. In conjunction, it will also cross-compare the real-world events with theoretical cyber scholars such as Kevin Mitnick and Bruce Schneier to illustrate the necessity for a holistic approach that prioritises human-centric cybersecurity strategies to ensure the preservation of cyber resilience.

Human errors

Cybersecurity is often perceived as a technological domain. However, as it governs and operates with people at the forefront, understanding the interplay between human psychology and cybersecurity vulnerabilities can lead to more effective risk mitigation strategies. Bruce Schneier (2018) argues that security is a process, not a product. In this regard, there is a need to emphasise that human behaviour and employee education must be continuously managed alongside technological advancements. Drawing on this

was the case study of the Equifax cybersecurity incident in September 2017, which occupies the top 10 charts of the largest data breaches in history with 148 million people affected (Kabanov, Ilya and Madnick, Stuart E., 2020). The case study reveals that the breach occurred due to an expired certificate on the Secure Sockets Layer (SSL) visibility appliance responsible for monitoring encrypted network traffic. The certificate expired in November 2016 and went unnoticed by the security team until August 2017. This oversight allowed an attack to commence in May 2017, resulting in the extraction of consumer personal identifiable information (PII) that went undetected for 78 days. The study highlights that while companies can integrate security technologies and develop high-fidelity detections, these measures are only effective if properly maintained and if poor or negligent human oversight is eliminated. In Verizon's 2024 Data Breach Investigations Report (2024), the company reported that approximately 68% of breaches involved a non-malicious human element, such as falling victim to social engineering or making an error. Remarkably, this was the same level as reported in 2020, indicating that in four years, no progress had been made to reduce the human element in security breaches.

Human psychology

Often attackers exploit human psychology rather than technical flaws to gain access to secure systems. The 2020 Twitter hack involved hackers using social engineering to manipulate employees into granting access to internal administrative tools (Witman & Mackelprang, 2022). This allowed them to take over high-profile accounts, including those of Barack Obama and Joe Biden, to promote a cryptocurrency scam. While this incident is not necessarily novel in its approach, it underscores the importance of addressing the risks posed by human psychology and the need for stronger internal security policies, such as multi-factor authentication and preventive education programmes to help reduce the likelihood of employees being exploited.

Despite significant advancements in cybersecurity technology and defensive strategies, human errors continue to be a major vulnerability

Former hacker and security consultant Kevin Mitnick illustrated how social engineering remains one of the most effective hacking techniques, exploiting emotions and the idea of trust - far easier and more effective than technical attacks - since human psychology can be fickle and not always equipped to handle situations in which individuals become 'hooked' (Mitnick in: Whiteman, 2017). This consideration highlights that, especially in this blind spot, organizations do not offer enough support to help their employees avoid emotional responses to attackers, leading many to act irrationally. Thus, the implementation of training programmes that heighten awareness of cyber-attacks can provide employees with more tools to stay vigilant (Stacey et al., 2021). A new model also focuses on using behavioural analysis to be monitored and trained with AI to detect anomalies in human behaviour and flag potential insider threats before they escalate (Verma, 2024). A comprehensive and multifaceted approach is required when addressing the complexity of human psychology. Enhancing security awareness training that incorporates phishing simulations and social engineering awareness techniques, alongside support to help employees manage their emotions during these simulations, will help strengthen this

blind spot by providing a more holistic combination of awareness and educational tools.

Lack of adaptation and investment

The current scope of cyber resilience primarily focuses on automation to account for the sheer volume and variety of threats. However, as examined above, most threats are human-correlated. Cavelty (2024) further contributes to this discourse by discussing the importance of investment within the economics of cybersecurity to address this primarily through financing improved security awareness and training programmes. When the healthcare sector became a prime target for cyber threats during the COVID-19 pandemic, phishing scams led to several breaches, such as the attack against the European Medicines Agency (EMA). The incidents often exploited pandemic-related anxieties, yet institutions like EMA, hospitals, and research institutions were not well trained to handle this new type of threat alongside the high emotional stakes of the pandemic situation (Awaludin et al., 2023). Psychological stressors such as these can lead to increased human susceptibility to phishing threats, and the lack of adaptation in cybersecurity training programmes during this time allowed attackers to gain access

> **“**
>
> **Often attackers exploit human psychology rather than technical flaws to gain access to secure systems**

to vaccine-related data through compromised emails and accounts (Awaludin et al., 2023). While brief, the analysis of attacks and cybersecurity in the health sector during the COVID-19 pandemic reaffirms the importance of organizations to adopt a multi-layered approach to mitigate human-related cybersecurity risks, combining both technological and psychological training.

Conclusion

The human element remains a significant challenge to cybersecurity resilience, and despite technological advancements, this area is still heavily neglected. As outlined in the case studies, such as the Equifax breach, Twitter Bitcoin scam, and EMA COVID-19 healthcare phishing attacks, major blind spots within the last five years are primarily social engineering and psychological vulnerabilities. Accordingly, the best methods to address this blind spot require further emphasis on education and training to improve awareness of social manipulation and psychological factors. Cybersecurity requires a holistic approach that prioritises education and human-centred strategies to reduce the current and evolving security risks.

References

Awaludin, A., Sulistyadi, W., & Chandra, A. F. (2023). Analysis of Attacks and Cybersecurity in the Health Sector During a Pandemic COVID-19: Scoping Review. Journal of Social Science, 4(1), 62-70.

Kabanov, Ilya and Madnick, Stuart E., A Systematic Study of the Control Failures in the Equifax Cybersecurity Incident (2020). MIT Sloan Research Paper No. 2020-19.

Cavelty, M. D. (2024). The Politics of Cyber-Security. Taylor & Francis.

Schneier, B. (2018). Click Here to Kill Everybody: Security and Survival in a Hyper-connected World. WW Norton & Company.

Stacey, P., Taylor, R., Olowosule, O., & Spanaki, K. (2021). Emotional Reactions and Coping Responses of Employees to a Cyber-attack: A Case Study. International Journal of Information Management, 58, 102298.

Verizon. (2020). 2020 Data Breach Investigations Report. Verizon Business.

Verizon. (2024). 2024 Data Breach Investigations Report. Verizon Business.

Verma, D. (2024). Enhancing Cybersecurity Through Adaptive Anomaly Detection Using Modern AI Techniques (Master's thesis).

Whiteman III, J. R. (2017). Social engineering: Humans Are the Prominent Reason for the Continuance of These Types of Attacks (Master's thesis, Utica College).

Witman, P. D., & Mackelprang, S. (2022). The 2020 Twitter Hack--So Many Lessons to Be Learned. Journal of Cybersecurity Education, Research and Practice, 2021(2).

ABOUT THE AUTHORS

Christopher Weir is a Technical Security Manager at Lloyds Banking Group with over seven years of experience in cybersecurity. Specializing in process automation, internal audits, and risk management, he leads initiatives to enhance the organization's security posture. Previously a Cyber Security Analyst, Christopher played a key role in identifying vulnerabilities and implementing solutions. His expertise in security policies, process automation, and auditing ensures the bank's systems remain secure and compliant with industry standards.

Dr Ally Zlatar is an artist, scholar, and activist. She is the founder of The Starving Artist, an artist initiative that utilises creative voices as a way to create advocacy and systemic reform since 2017. She has received numerous accolades for her humanitarian work, such as the Commonwealth Innovation Award (2023), UN Women 30 for 2030 (2024), the Princess Diana Legacy Award (2021), and the King Hamad Award for Youth Empowerment (2022).

Download UNICRI publications

“

**In a few clicks,
a fake video of a politician
or a cloned voice of a CEO
can spread chaos,
from election meddling
to million-dollar scams**

Escaping Plato's digital cave: deepfakes, cybersecurity, and the battle for truth

by Leonardo Lazzaro

Truth feels slippery in today's digital world. Deepfakes - videos, voices, or images crafted by artificial intelligence (AI) to look and sound strikingly real - are tearing at the fabric of what we believe. They do not just disrupt how we communicate; they threaten our dignity, unravel democratic trust, and shake the foundations of shared knowledge.[1] In a few clicks, a fake video of a politician or a cloned voice of a CEO can spread chaos, from election meddling to million-dollar scams.

This is not a new story, though it wears a high-tech mask. Plato's allegory of the cave, where prisoners mistake shadows for reality, feels like it was written for us. Our cave is the Internet, and deepfakes are the shadows, algorithmic illusions that trick us into believing falsehoods.[2] Back then, shadows came from firelight; now, they are born from AI code. This leap from mere misrepresentation to outright fabrication undermines the reliability of what we see and hear, turning photos, videos, and voices into potential lies.[3]

Cybersecurity experts see deepfakes as a runaway train: they power disinformation campaigns that sway voters, as seen in 2024's flood of fake election

1 Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review.

2 Floridi, L. (2021). The logic of information: A theory of philosophy as conceptual design. Oxford University Press.

3 Kietzmann, J. (2019). Deepfakes: Trick or treat? Business Horizons.

videos. They fuel harassment, extortion, and identity theft, targeting mostly women and girls with non-consensual deepfakes of sexual nature that violate privacy and dignity.[4] In finance and healthcare, fake records or identities can wreak havoc, undermining trust and security. Deepfakes thrive because they exploit how we trust our senses, spreading like wildfire on social media. They create an epistemic crisis, forcing us to rethink what we accept as real.[5]

The European Union (EU) is fighting back. The Network and Information Systems (NIS) 2 Directive,[6] rolled out in October 2024, pushes Member States to raise awareness and teach cyber hygiene, zeroing in on deepfakes as a clear and present danger. In this context, Italy's leading the charge with a 2024 draft law that forces digital platforms to label deepfake content or face penalties. The new article 612-quater in the Criminal Code criminalize the spreading of harmful AI-altered media, punishable by one to five years in prison.

But laws and firewalls cannot win this alone. Deepfakes hack our minds, playing on emotions and instincts. The real battle is cultural, and young people are our best hope. Teens and young adults are not just glued to their screens; they are creating and sharing content, shaping the digital world. Teaching them to spot and call out fake media is non-negotiable.[7] Programmes like the EU's Cybersecurity Skills Academy and United Nations' (UN) digital literacy initiatives are stepping up, turning youth into guardians of truth.[8] They are not just users; they are builders of a digital culture that values honesty and fairness. Universities are also retooling their courses, blending cyber law, AI ethics, and data governance to train the next generation of tech-savvy jurists.

Data protection is another key weapon. Deepfakes feed on personal data, photos, voice clips, or videos scraped from social media without permission. The General Data Protection Regulation (GDPR) requirements, like keeping data use minimal and lawful, can hinder these technologies by limiting unauthorized access to our digital selves.[9] NIS 2 builds on this, demanding accountability from everyone handling data, not just to stop breaches but to prevent ethical misuse. It is about treating data as a treasure to guard, not just a box to check for compliance.[10]

The old adage "seeing is believing" no longer holds. Trust the glue that holds our societies together, is under siege

4 Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.

5 European Union Agency for Cybersecurity. (2023). Threat landscape 2023.

6 Regulation (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive). Official Journal of the European Union.

7 UNICEF. (2019). Digital literacy for children and young people: An essential component of digital citizenship.

8 European Commission. (2023). Cybersecurity skills academy.

9 González Fuster, G. (2014). The emergence of personal data protection as a fundamental right of the EU. Springer.

10 Floridi, L., & Cowls, J. (2019). *A unified framework of five principles for AI in society.*

The old adage "seeing is believing" no longer holds. Trust, the glue that holds our societies together, is under siege.[11] Cybersecurity is no longer just the domain of IT departments; it has become a collective responsibility, bringing together legal frameworks, technological tools, civic engagement, and ethical awareness. Critical thinking is our first line of defence against the persuasive power of deepfakes. Sometimes, all it takes is one misstep - a careless click, a weak password - to trigger a chain reaction of chaos, exposing entire organizations to risk. NIS 2 recognises this reality, demanding that digital security become a shared duty, from interns to executives. Resilience is not a top-down solution, it must be embedded across all levels of our workplaces and communities, making cybersecurity a true team effort. Deepfakes are not just a tech problem; they are a wake-up call for how we think and talk in public. They tamper with honest conversations and real voices. Beating them takes more than staying alert, it demands a cultural shift. If Plato's cave is our digital world, young people are our guides to the light.

> **"If Plato's cave is our digital world, young people are our guides to the light**

By teaching them to question what they see and empowering their innovations, we can build a digital space that defends truth, protects justice, and lives up to UN Sustainable Development Goal (SDG) 16's promise of strong and fair institutions. Let us back them with policies that fund their tools, weave their ideas into national plans, and train them to lead. Only then can we escape the shadows and reclaim reality.

ABOUT THE AUTHOR

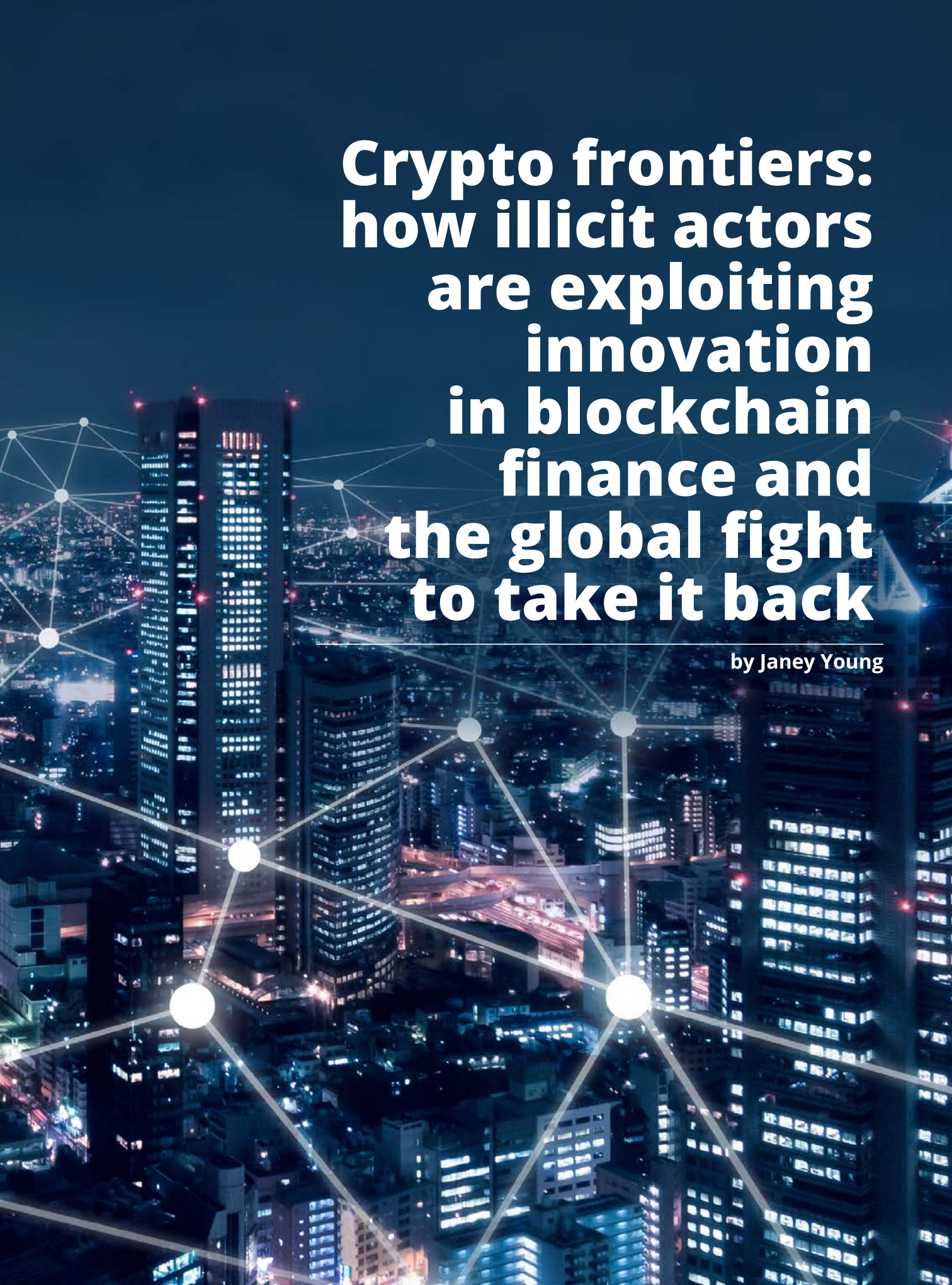
Leonardo Lazzaro is a trainee lawyer at Baker McKenzie Italy in the Commercial, Data and International Trade practice, where he focuses on privacy, cybersecurity, and Intellectual Property law. He recently completed his law degree at Luiss Guido Carli University, graduating with 110 summa cum laude and Special Acknowledgment for his thesis on the evolving role of consent as a point of tension between data protection and market regulation. He also deepened his knowledge of these legal domains during an academic experience in the United States at Fordham University.

11 OECD. (2019). Recommendation of the Council on Artificial Intelligence.

“Plato’s allegory of the cave, where prisoners mistake shadows for reality, feels like it was written for us. Our cave is the Internet, and deepfakes are the shadows, algorithmic illusions that trick us into believing falsehoods

Crypto frontiers: how illicit actors are exploiting innovation in blockchain finance and the global fight to take it back

by Janey Young



The rise of blockchain technology has unlocked unprecedented opportunities in global finance. What began with Bitcoin has now evolved into a diverse digital ecosystem powered by decentralized networks like Ethereum, Polygon, Solana, and others. These platforms have laid the foundation for a wave of innovation — from decentralized finance (DeFi) to non-fungible tokens (NFTs) — promising more accessible, transparent, and efficient financial systems.

But as with many transformative technologies, the same features that drive progress can also invite abuse. Criminal groups are rapidly adapting, exploiting blockchain's decentralized and pseudonymous architecture to commit fraud, launder money, and evade law enforcement — presenting new and complex challenges to global security and governance.

Decentralization's double-edged sword

At its core, blockchain offers transparency and immutability. Every transaction is recorded on a public ledger, visible to anyone with Internet access. However, the decentralization that makes blockchain powerful also makes it difficult to regulate. Without intermediaries like banks or payment processors, there are fewer gatekeepers to monitor activity or flag suspicious behaviour. This appeals not only to technology enthusiasts but also to money launderers, cybercriminals, and fraudsters.

In particular, DeFi — blockchain-based financial services that operate without intermediaries and enable activities such as trading, borrowing, and lending, among others — has emerged as fertile ground for exploitation. Many operate without full know-your-customer (KYC) verifications, allowing users to interact anonymously or pseudonymously. The result is an ecosystem where significant funds can move at pace and across borders with limited oversight.

Vulnerabilities in decentralized finance: innovation meets exploitation

While DeFi's open-source ethos drives innovation, it also creates serious vulnerabilities. Weaknesses include poorly audited smart contracts, flawed governance mechanisms, and the manipulation of price oracles and liquidity pools. The ability to launch rapid, high-impact attacks has become a defining risk within the DeFi ecosystem. Exploits such as reentrancy attacks — where funds are repeatedly withdrawn from a smart contract before balances can update — or the distortion of data feeds to falsify asset

values illustrate the systemic risks. Inadequately secured platforms further expose users to malicious actors who may seize control of operations, leading to financial losses, service disruptions, or outright theft.

Among the most concerning methods exploiting these vulnerabilities are flash loan attacks. Flash loans enable users to borrow vast sums without collateral, provided the loan is repaid within the same transaction. Because these attacks can unfold in seconds, bypassing conventional fraud detection, criminals have used them to manipulate asset prices or drain liquidity pools. In some instances, attackers have even leveraged flash loans to gain temporary control of governance tokens, allowing them to force through malicious proposals — such as transferring treasury funds to their own wallets — before the broader community can respond.

Non-fungible tokens: digital art or digital laundering?

NFTs have captured popular imagination as digital collectibles and artworks, from meme coins to celebrity-endorsed avatars. But beneath the surface, they also present new opportunities for abuse. Criminals can mint NFTs, then engage in 'wash trading' — selling them to themselves using illicit funds. They can then offload them to unsuspecting buyers, effectively laundering money through digital art. This risk is compounded as many NFT marketplaces lack strong identity verification or compliance mechanisms.

Furthermore, 'rug pulls' have become a recurring threat, where developers of 'meme' tokens — a type of cryptocurrency created largely as a joke or Internet trend, whose value is driven by online hype rather than real-world use — vanish after securing investor funds. A high-profile example is the 2021 Squid Game (SQUID) token scam, in which the creators drained all liquidity and disappeared, leaving

investors with worthless tokens.[1]

Obscuring the trail: mixers, tumblers, and privacy coins

To further mask their tracks, illicit actors often turn to transaction-obfuscating services known as mixers or tumblers. These tools blend different users' transactions, making it harder to trace individual fund flows. Others rely on privacy-focused cryptocurrencies like Monero, which are designed to resist tracking altogether. While all these tools have legitimate privacy use cases, they are increasingly flagged by regulators as high-risk, and several cryptocurrency mixers have been sanctioned or dismantled in global efforts to stem illicit activity.

The myth of the untraceable blockchain

Despite these challenges, the belief that crypto transactions are untraceable is a misconception. Public blockchains, by design, maintain a transparent record of all activity, allowing investigators to follow the digital money trail with growing sophistication.

> **Despite these challenges, the belief that crypto transactions are untraceable is a misconception**

Blockchain forensics tools have become essential in modern law enforcement, enabling agencies to trace complex transaction chains across networks. As these tools evolve, they are helping expose criminal networks and recover stolen funds.

A united front: international and industry collaboration

Addressing the dark side of crypto requires global coordination. Initiatives such as the Egmont Group's information-sharing among Financial Intelligence Units (FIUs)[2] and INTERPOL's Global Complex for Innovation[3] are strengthening cross-border efforts to investigate and disrupt criminal activity. Meanwhile, the Financial Action Task Force (FATF) continues to refine its guidance on virtual assets, urging Member States to adopt clear and robust regulatory frameworks.[4] In response, many jurisdictions are integrating these standards into national legislation, clarifying compliance obligations and boosting their capacity to monitor digital financial crime. Building on these efforts, the private sector plays a vital role in countering crypto crime. Blockchain analytics firms are evolving their technologies to keep pace with threats, deploying tools to trace illicit transactions across decentralized and often opaque networks. Through close cooperation with law enforcement, cryptocurrency exchanges — platforms where digital assets are bought, sold, and traded — and other industry stakeholders are helping to secure vulnerabilities and identify, investigate, and disrupt suspicious activity in real time. Furthermore, many exchanges are now proactively freezing wallets tied to criminal activity and reinforcing compliance to align with international standards, contributing to a more secure digital asset ecosystem.

1 Chris Stokel-Walker, "How a Squid Game Crypto Scam Got Away With Millions", Wired, November 2021.
2 "The Egmont Group of Financial Intelligence Units", Financial Crimes Enforcement Network.
3 "INTERPOL Innovation Centre".
4 "FATF Urges Stronger Global Action to Address Illicit Finance Risks in Virtual Assets", FATF, June 2025.

Striking the balance: innovation without impunity

Policymakers face a delicate balancing act: how to foster innovation without opening the floodgates to exploitation. Overregulation could stifle progress and push activity underground. Underregulation and criminal misuse become inevitable.

Ongoing dialogue is critical between governments, industry leaders, technologists, and civil society. So is public education. Many users remain unaware of the risks associated with DeFi platforms and NFTs. Promoting digital literacy, responsible platform design, and stronger consumer protection can empower individuals and reduce the impact of fraud.

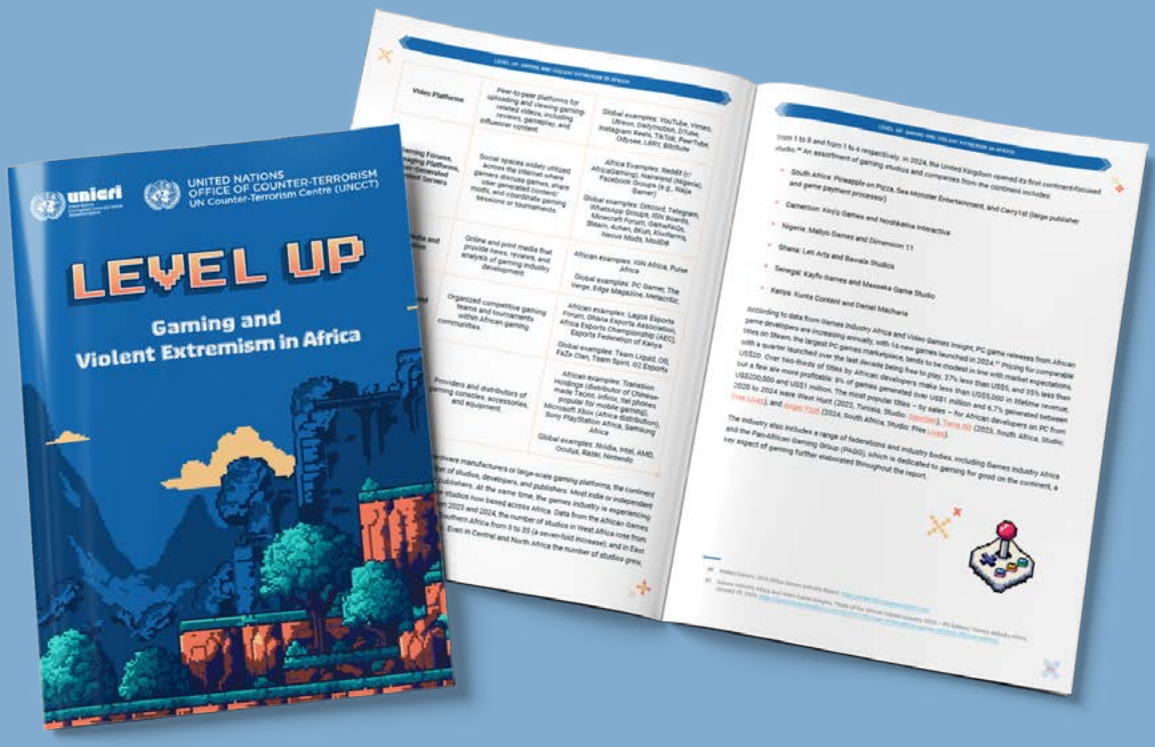
Looking forward: a safer digital finance future

The frontier of blockchain finance is expanding rapidly, and so is its misuse. As decentralized systems continue to reshape global finance, the international community must act decisively to secure this space.

The international community must stay ahead of the curve. Through agile regulation, shared intelligence, and continued investment in investigative tools, we can disrupt illicit activity while preserving the transformative potential of blockchain. Only then can we ensure that this technology serves the cause of global peace, economic resilience, and inclusive development.

ABOUT THE AUTHOR

Janey Young is the Founder and Principal Consultant of Safe Digital Futures, advising organizations on cybercrime, risk management, and responsible innovation. Her career includes leading cybercrime investigations at the UK's National Crime Agency and Europol's European Cybercrime Centre, where she pioneered strategies against threats like NotPetya and dark web activity. Most recently, as Head of Global Investigations at Chainalysis, she built internationally recognized capabilities using blockchain, AI, and advanced analytics to strengthen compliance and investigative solutions worldwide.



Download UNICRI publications

Gaming the system: closing anti-money laundering gaps in the digital entertainment economy

by Adam Rousselle and Galen Lamphere-Englund

Gaming platforms now operate as parallel financial systems. What began as microtransactions and virtual rewards has grown into a multibillion-dollar marketplace where users buy, sell, and trade assets with real monetary value. These flows create exchangeable currencies, digital wallets, and peer-to-peer markets that criminals can exploit to move money across borders with limited oversight.[1]

This process creates a laundering loop that hides the origin of funds while blending them into massive, legitimate transaction volumes

Money launderers, organized criminals, terrorists, and violent extremist organizations (VEOs) increasingly test these tools. They convert in-game tokens into fiat or cryptocurrency, route funds through streaming donations, and exploit loopholes in marketplace policies. Regulators, however, remain slow to respond. The Financial Action Task Force (FATF) first warned about these risks in 2018, but most gaming platforms still operate without meaningful anti-money laundering (AML) obligations.[2]

Emerging typologies in gaming and streaming

Criminals exploit gaming ecosystems because they offer liquidity, scale, and weak oversight. Several typologies already show how illicit finance flows through these platforms.

In-game currency laundering. Platforms like Roblox and Fortnite generate nearly $10 billion in annual revenue through virtual currencies such as Robux[3] and V-Bucks.[4] Criminals can buy tokens with illicit funds, trade them inside the game, and then cash out through developers or secondary markets.[5] This process creates a laundering loop that hides the origin of funds while blending them into massive, legitimate transaction volumes.

Loot box and in-game item exploitation. Loot boxes — randomized purchased rewards often delivered as in-game items — let users assign monetary value to digital assets that regulators cannot easily trace. Launderers purchase loot boxes with illicit funds, resell high-value items on external marketplaces, and convert the proceeds back into fiat or cryptocurrency.[6] The opacity of these probability-based rewards makes it difficult for authorities to track flows.[7]

Third-party currency conversion. A growing industry of external exchanges converts both "non-convertible" and "convertible" in-game currencies into cash or crypto.[8] These services bypass platform-level controls — operating in regulatory gray zones — enabling criminals to arbitrage value cross-jurisdictionally.[9]

Blockchain-enabled livestreaming. Streaming platforms like DLive integrate crypto payment rails directly into content ecosystems. Even after crackdowns on mainstream ser-

1 "Money in Video Games: It's Virtually Everywhere!", *Northbrook Bank*, 2023.
2 *FATF*, "Guidance for a Risk-Based Approach to Virtual Assets and Virtual Asset Service Providers", 2019.
3 Moshe Klein, "Video Games Might Matter for Terrorist Financing", *Lawfare Media,* May 2024.
4 Greg Belding, "In-game Currency & Money Laundering Schemes: Fortnite, World of Warcraft & More", *Infosec*, June 2021.
5 Skye Jacobs, "Microtransactions Accounted for 58% of PC Gaming Revenue Last Year", *TechSpot,* April 2025.
6 Mavis Bennett, "The Potential Perils of Online Gaming", *Acams Today,* June 2023.
7 Leon Y. Xiao, "Regulating Loot Boxes as Gambling? Towards a Combined Legal and Self-Regulatory Consumer Protection Approach", *Interactive Entertainment Law Review*, June 2021.
8 *Consumer Financial Protection Bureau*, "Banking in Video Games and Virtual Worlds", April 2024.
9 Shane Kelly, "Money Laundering through Virtual Worlds of Video Games: Recommendations for a New Approach to AMLRegulation", *Syracuse University College of Law*, 2022.

vices, these ecosystems allow users to solicit donations and monetize propaganda with little oversight, giving banned actors a resilient financial lifeline.[10]

Violent extremist experimentation. Some VEOs have begun experimenting with non-fungible tokens (NFTs), currency, and crypto collectibles hosted on blockchain platforms.[11] While not yet actualized, these efforts suggest growing adaptation to new financetech mediums. Given that VEOs tend to probe early and scale later, this experimentation could represent an emerging threat vector.

Collectively, these typologies show why gaming economies matter in the global finance landscape. They blend entertainment, finance, and social media in ways that regulators still treat as peripheral.

In practice, they function as global value-transfer systems with billions in legitimate flows, the kind of environment illicit actors seek to exploit.
By understanding these typologies, we can better understand global regulatory gaps and how to address them.

10 Ariel Bogle, "Buying and Selling Extremism: New Funding Opportunities in the Right-Wing Extremist Online Ecosystem", *ASPI*, 2021; Ben Makuch, "Trump's Promise to Loosen Crypto Regulations May Be Boon for Extremist Groups", *The Guardian*, November 2024.

11 Mahmoud Firas, "The Gamification of Jihad: Playing with Religion", *Danish Institute for International Studies*, 2022; Julia Handle and Louis Jarvers, "Extremist NFTs Across Blockchains", *Lawfare Media*, May 2023.

Regulatory landscape and gaps

Governments have begun sketching frameworks for digital finance, but gaming and streaming ecosystems remain largely unaddressed. The Algeria Principles, adopted by the UN Security Council in 2025, call for better understanding of terrorist financing risks in emerging technologies and for proportionate regulation.[12] In 2018, FATF added Recommendation 15, urging states to regulate virtual assets and services tied to new technologies.[13] Yet neither framework translates directly into gaming.

The picture remains uneven across jurisdictions. The European Union's Digital Services Act and Terrorist Content Online rules impose broad content obligations but leave gaming platforms in a gray zone.[14] The United Kingdom and Australia emphasize online safety and child protection via Ofcom and the E-Safety Commissioner, not anti-money laundering per se.[15] In the United States, proposals like the Children's Online Privacy Protection Act (COPPA) 2.0 point to greater accountability, but stop short of imposing AML standards on gaming-related flows.[16] Each region touches a piece of the problem, yet none imposes consistent financial scrutiny.

> **Gaming companies that handle billions in microtransactions have no universal AML obligations, even though their systems function like financial intermediaries**

These gaps converge into structural blind spots. Gaming companies that handle billions in microtransactions have no universal AML obligations, even though their systems function like financial intermediaries. Jurisdictional arbitrage lets bad actors exploit inconsistencies, moving value through platforms headquartered in permissive environments. Crowdfunding and tipping systems, particularly on livestreaming services, further widen the gap: users can convert attention directly into money with little oversight or traceability.

The result is a patchwork of digital safety laws that largely ignore the financial threats posed by in-game economies. This mismatch leaves governments chasing symptoms rather than building coherent oversight.

Ireland's strategic role

Ireland occupies a unique position in this landscape. As the European headquarters for Apple, Google, and many leading game publishers, Ireland processes much of the global revenue from in-game purchases and digital wallet flows. That puts Dublin at the nexus of streaming, gaming, and fintech flows — making its regulatory posture internationally critical to how platforms manage AML, wallet conversions, and microtransaction rails.

In recent years, the Central Bank of Ireland has expanded its supervision of Virtual Asset Service Providers (VASPs) and pushed for stronger anti-money laundering measures across fintech ecosystems.[17] These frameworks extend pressure onto companies routing

12 UN Security Council – Counter Terrorism Committee, "Algeria Guiding Principles".

13 FATF, "Public Statement on Virtual Assets and Related Providers", 2019.

14 European Commission, "Terrorist Content Online".

15 Jonathan Frost, "Fraud and the UK's Online Safety Act", *BioCatch; E-Safety Commissioner*, "Online Safety Act".

16 "COPPA 2.0 Reintroduced – What You Need to Know", *BBB National Programs*, May 2025.

17 Central Bank of Ireland, "Central Bank Highlights Weaknesses in Virtual Asset Service Providers' AML/CFT Frameworks", July 2022; Jennifer Flannery, "Unveiling the Financial Crime/Anti-Money Laundering (AML) Dynamics of Ireland's Payments Sector", *Grand Thornton*, May 2024.

gaming transactions through Irish subsidiaries. That leverage matters. Because item and loot box sales, microtransactions, and streaming monetization often pass through Irish-registered entities, decisions made in Dublin ripple globally. By tightening AML standards, Ireland can set de facto norms for the industry and force multinational firms to apply compliance standards worldwide. Although not the largest regulator, Ireland's jurisdictional rules can reshape how digital economies handle illicit finance far beyond its borders.

Conclusion: convergence risks

Gaming and streaming ecosystems are no longer fringe — they are now embedded in the wider architecture of global financial flows. With billions moving through microtransactions, gift economies, and crypto-linked rails, these platforms represent a growing vector for illicit liquidity. Absent harmonized oversight, jurisdictional gaps will continue to be exploited at scale, enabling threat actors to launder value as easily as they share content, making systemic exploitation inevitable. Ireland's expanding VASP supervision highlights the role that forward-leaning jurisdictions can play, but systemic resilience will require coordinated action from regulators, platforms, and financial authorities working in tandem.

ABOUT THE AUTHORS

Adam Rousselle is an analyst and advisor specialising in threat finance. His work has been cited by the Financial Action Task Force (FATF), the United Nations (UN), and multiple government reports. He is the founder of Between the Lines Research and advises on illicit-finance issues internationally.

Galen Lamphere-Englund is an international security advisor specialising in counter-terrorism and technology-driven online harms. He has delivered research and advisory work for UN agencies, governments, NGOs, and major tech firms. He co-founded the Extremism and Gaming Research Network and serves as team lead at the Christchurch Call Foundation.

Criminals exploit gaming ecosystems because they offer liquidity, scale, and weak oversight

Using cyberweapons to steal trade secrets

by Vasilis Katos, Kenneth Wright and John Zacharia

Introduction

Trade secrets are among the most valuable forms of intellectual property. Increasingly, like other valuable proprietary information, trade secrets are stored in digital form. As with any information stored digitally, trade secrets are vulnerable to cyberattacks. This article explores the definition of trade secrets, examines the cyberweapons online criminals use to exfiltrate digital information, and discusses how such cyberweapons may be used to steal trade secrets.

Trade secrets

Description of trade secrets

In today's knowledge-driven economy, the value of information can easily eclipse that of physical assets. Among the most critical yet least visible forms of intellectual property (IP) are trade secrets, confidential business information that offers companies a strategic advantage precisely because it remains undisclosed to competitors.

Trade secrets span a wide variety of proprietary knowledge, ranging from customer databases and pricing strategies to complex algorithms and chemical formulas. To qualify as a trade secret, information must meet three core criteria. First, it must be secret, meaning it is not generally known or readily accessible to people who typically deal with such information (competitors). Second, it must have commercial value to competitors. And third, the business must take reasonable steps to safeguard its confidentiality, such as implementing non-disclosure agreements, limiting physical or digital access, and training employees on information-security practices.

> **"As businesses continue to operate in an environment of rapid innovation and high employee mobility, the strategic management of trade secrets is becoming as essential as protecting patents or trademarks.**

The legal frameworks governing trade secrets vary globally but share foundational principles. In the United States, trade secrets are protected under both State (mostly based on the Uniform Trade Secrets Act (UTSA) and Federal law pursuant to the Economic Espionage Act (EEA). The EEA defines trade secrets broadly as all forms of tangible or intangible information that derive economic value from not being generally known to competitors and are subject to reasonable confidentiality measures.

The European Union (EU) introduced a harmonised approach with the EU Trade Secrets Directive (Directive (EU) 2016/943). This legislation aligns Member

States around a common definition, stating that a trade secret is information that is secret, has commercial value due to its secrecy, and has been reasonably protected by its holder. The Directive aims to balance business protection with transparency and the mobility of workers.

Unlike patents, trade secrets can offer potentially indefinite protection, but only as long as the information remains confidential. The risk, however, is that once exposed, a trade secret loses its legal shield. Misappropriation – acquiring

> **“**
>
> **In today’s knowledge-driven economy, the value of information can easily eclipse that of physical assets. Among the most critical yet least visible forms of intellectual property (IP) are trade secrets**

or using trade secrets through improper means such as theft, bribery, or breach of contract – can lead to significant civil and, in some cases, criminal liability. Civil remedies typically include injunctions, damages, and in urgent cases, seizure of stolen materials under court order. Criminal remedies include imprisonment.

As businesses continue to operate in an environment of rapid innovation and high employee mobility, the strategic management of trade secrets is becoming as essential as protecting patents or trademarks. Organizations that systematically identify, secure, and monitor their proprietary information stand to benefit not only from legal safeguards but from

a sustainable competitive edge – one that remains out of sight, but very much in force.

Description of types of trade secrets that are typically stored in digital form on a server (e.g., prototype design of a new semiconductor chip, pre-release music/movie/ fashion design)

As innovation moves increasingly into digital environments, trade secrets, once only locked in filing cabinets, are now stored on servers and cloud platforms. These high-value assets span industries and are central to competitive advantage, yet they are also highly vulnerable to cyber threats and internal leaks.

In the technology and manufacturing sectors, prototype designs such as digital files for semiconductor chips or consumer electronics are prime targets. Their exposure can erode competitive positioning and disrupt product roadmaps. Similarly, pre-release content in media – like unreleased films, music, or video games – is closely guarded to prevent leaks that could damage revenue or marketing strategies.

Source code, proprietary algorithms, and training data are the crown jewels of software and artificial intelligence (AI) firms. Stored in private repositories, they represent years of research and development and are often at the core of a company's IP portfolio. In the life sciences and chemicals industries, digital product formulations and test data are essential trade secrets, carefully managed through secured data environments.

Even creative sectors such as fashion and architecture rely on digital protection for pre-release designs and campaign assets, which can be compromised long before public launch. Alongside these are internal strategic documents – pricing models, product roadmaps, or market entry plans – which, if exposed, can undermine competitive moves.

As more of these assets are stored and exchanged in digital form, robust cybersecurity, access control, and a strong culture of confidentiality are essential. In an age where information moves fast and threats evolve rapidly, safeguarding digital trade secrets is no longer optional – it is foundational.

Economic impact if trade secrets are stolen

Trade secrets are often the invisible backbone of a company's competitive edge. When these secrets are stolen, the economic fallout can be devastating, not just for individual businesses but for entire industries and economies.

At the corporate level, theft of trade secrets can rapidly erode a company's market position. Competitors armed with proprietary formulas, designs, or strategies can flood the market with knockoffs or improved products at lower prices, undercutting the original innovator. This leads to lost sales, declining profits, and in some cases, irreversible brand damage. The financial blow does not stop there - companies face substantial costs related to legal battles, cybersecurity upgrades, and crisis manage-

ment efforts to mitigate the damage.

Investor confidence can also take a hit when sensitive information is compromised. A breach signals potential vulnerabilities in management and governance, often triggering stock price drops and reduced capital inflows. In fast-paced industries like technology or pharmaceuticals, where first-mover advantage is critical, losing trade secrets can translate into lost opportunities worth millions or even billions of dollars.

On a broader scale, the theft of trade secrets poses a significant threat to national economies. Cyber-espionage campaigns targeting proprietary innovations disrupt innovation cycles and discourage research and development investments. This erosion of intellectual property protection undermines the very incentives that fuel economic growth and technological advancement. Furthermore, it can lead to job losses as companies scale back or relocate operations to safer jurisdictions.

> **"As innovation moves increasingly into digital environments, trade secrets, once only locked in filing cabinets, are now stored on servers and cloud platforms**

Governments and industry leaders increasingly recognise that safeguarding trade secrets is essential not only for individual business success but for maintaining global economic competitiveness. The stakes are high, and the economic impact of stolen trade secrets is a clarion call for stronger protections and proactive defence strategies.

Cybercrime

Cyberweapons used today

The dominant 'ingredient' present in a trade secret theft attack is data exfiltration. Unlike attacks such as ransomware, or Denial of Service, an attacker targeting a company's trade secrets aims to be undetected throughout and after the attack. This profile of an attack is akin to the operations and behaviour exhibited by Advanced Persistent Threats (APTs). Another significant risk is from insider threats, namely employees, partners, and third parties in general who already have legitimate access to the organization's systems.

How AI supercharged these cyberweapons

AI has significantly amplified the sophistication and effectiveness of trade secret theft attacks, particularly through enhanced spear phishing capabilities. AI-powered language models can now generate highly convincing, personalised phishing emails that mimic writing styles, reference specific industry terminology, and incorporate contextual details gathered from social media and public sources. Machine learning algorithms can analyse vast amounts of data to identify optimal targets, timing, and attack vectors, while automated systems can conduct reconnaissance at unprecedented scale and speed. AI also enables dynamic attack adaptation, where malicious systems can modify their behaviour in real-time based on

> **AI has significantly amplified the sophistication and effectiveness of trade secret theft attacks, particularly through enhanced spear phishing capabilities**

defensive responses, making detection more challenging. Furthermore, deepfake technology and voice synthesis can create convincing audio or video content for social engineering attacks, while AI-driven automation allows attackers to manage multiple simultaneous campaigns against different targets. Perhaps most concerning is AI's ability to help attackers blend malicious activities with normal network behaviour patterns, making the stealthy, long-term data exfiltration characteristic of APTs even more difficult to detect through traditional security monitoring systems.

How cyberweapons can be/have been used to steal trade secrets

Hacking a server to obtain a trade secret

Trade secrets in digital form, such as the formula for a breakthrough drug that has not yet been patented, are typically stored on a server that is armed with firewalls, encryption, and data logging. Trade secret owners use these tools to limit who has access, restrict the means of access, and to track who accessed the trade secrets and how much data was accessed at any given time.

A trade secret thief may use AI-enabled machine learning and adversarial neural networks to attempt to detect and decode the encryption used by the trade secret owner and circumvent the server's firewall. Once breached, the thief may then deploy AI-enabled APTs into the owner's network to wait until the data traffic on the server peaks so as to reduce the likelihood of detection during the exfiltration process of the trade secret. Using these cyberweapons, the thief could steal the aforementioned drug formula before its true owner had the opportunity to patent and market the drug.

AI-enabled phishing emails/communications

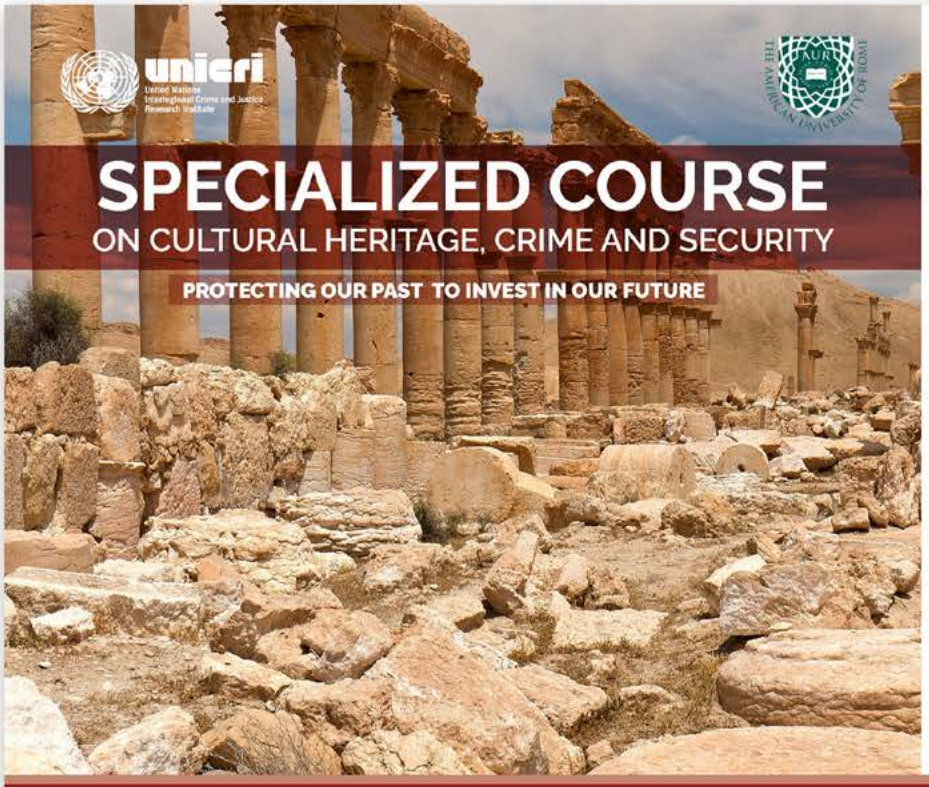
Some trade secrets, such as pre-release music albums or movies, are stored on servers with access limited to only a small team of employees. To identify the weakest employee with access to the trade secret, a trade secret thief may use AI-enabled automated systems to conduct rapid reconnaissance and determine to whom a phishing email should be directed. Once that employee is identified, the thief can implement AI-enabled deepfake technology to imitate the voice of the CEO of the company instructing the employee to disclose the trade secret via email. In this scenario, the thief could use the same AI-enabled cyberweapons referenced in the previous scenario to hack into the CEO's email account. Once the duped employee sends the email attaching the trade secret, the thief unlawfully accesses the CEO's email account (without authorisation) and downloads the trade secret. In this way, the thief

ABOUT THE AUTHORS

Kenneth Wright has a MSc in Information Security and previously served as a senior police officer but for the last 26 years has worked with governments, law enforcement and the private sector - in over 70 countries - to protect and enforce intellectual property (IP). During this period, he has prepared national IP strategies, drafted IP legislation, delivered IP capacity building and investigated IP crimes.

Vasilis Katos is Professor of Cyber Security and lead of Bournemouth University's Computer Security and Incident Response Team, BU-CERT. Prof. Katos specializes in digital forensics for intellectual property crime investigations. His work focuses on applying cybercrime and cybersecurity incident response practices to online investigations of IP crime.

John H. Zacharia is the Founder of Zacharia Law PLLC, a law firm dedicated to helping clients combat cyber theft and protect and enforce their intellectual property rights. Previously, as the Assistant Deputy Chief for Litigation of the Computer Crime and Intellectual Property Section (CCIPS) of the United States Department of Justice's Criminal Division, John was responsible for supervising all of the intellectual property and cybercrime prosecutions by the Section's 40 attorneys. While at CCIPS, John became one of the most experienced federal prosecutors of intellectual property crime in the country.



unicri
United Nations Interregional Crime and Justice Research Institute

SPECIALIZED COURSE
ON CULTURAL HERITAGE, CRIME AND SECURITY

PROTECTING OUR PAST TO INVEST IN OUR FUTURE

Join

the **United Nations Interregional Crime and Justice Research Institute** and **The American University of Rome (AUR)** for an in-depth program to safeguard cultural heritage in a time of crisis.

Explore the legal, political, and criminal challenges of protecting cultural property amidst conflict, looting and trafficking.

Info

- **Expert insights.**
- Practical **exercises** and **simulations.**
- Real-world **case studies.**
- Certificate of Participation issued by **UNICRI** and **AUR.**

The online format offers flexibility, bringing together participants from around the world for a truly global classroom experience.

English | Online | Students, Post-Graduates and Professionals

Be part of the global movement to safeguard cultural heritage!

unicri
United Nations Interregional Crime and Justice Research Institute

SIOI
UNA Italy

SUMMER SCHOOL
ON MISINFORMATION, DISINFORMATION AND HATE SPEECH

Join

the **United Nations Interregional Crime and Justice Research Institute** and the **Italian Society for International Organization (SIOI)** for a dynamic hybrid program to safeguard truth in the digital age.

Address the legal, political, technological, and ethical challenges of misinformation, disinformation and hate speech.

Info

- **Expert insights.**
- Practical **exercises** and **simulations.**
- Real-world **case studies.**
- Certificate of Participation issued by **UNICRI** and **SIOI.**

This hybrid program allows for all students to have an immersive and adaptable experience, in a truly global classroom.

English | Hybrid (Rome or Online) | Students, Post-Graduates and Professionals

Take action to safeguard the digital information ecosystem. Be part of the solution!

Every digital interaction runs on infrastructure consuming vast amounts of electricity and water; a single hyperscale data center can use more than 1.5 million gallons daily

From stolen data to stolen resources

by Phoenix Omwando

The most dangerous cyber threats today do not look like threats at all and most of us will not see them coming until it is too late. They hide in job offers that seem legitimate, in conversations that feel private, in technologies that promise to change the world. They do not crash your system or lock your files. Instead, they take something far more valuable: pieces of you that you can never get back. And by the time you realize what has happened, your data is already somewhere you will never be able to reach.

I learned this the hard way. My personal email address felt like a drop in the ocean — small, invisible and unimportant. That illusion shattered the day I began receiving emails from my own address, threatening to leak fabricated videos unless I paid in Bitcoin. They even included a photo of my home. Seeing my own name and house made my stomach drop. For weeks afterward, my inbox filled with login alerts from dozens of countries. My first instinct was to dismiss it as crude spam. But this wasn't just noise. It was a symptom of something bigger. My personal data had been taken, traded, and turned into currency. Once stolen, it was not sitting on one server waiting to be deleted. It was circulating through an invisible economy, bought and sold by people I would never meet, in countries I might never visit. Until that moment, I had only a vague idea of how personal and irreversible data theft could be – but I learned almost overnight just how real it was.

> **“The most dangerous cyber threats today do not look like threats at all and most of us will not see them coming until it is too late**

That is the reality of the digital landscape today. While some cyberattacks make headlines by disrupting critical systems, the most insidious threats often operate quietly, targeting individuals in ways that are only noticed after the fact. They

wear the face of normality, slipping past suspicion until it is too late. And one of the fastest-growing examples is happening in a place we are taught to trust: the job market.

Across Europe, Southeast Asia, and Africa, fully fledged operations now pose as legitimate companies. INTERPOL's 2023 cybercrime report found that recruitment scams targeting young job seekers increased by more than 30% in a single year. In the EU alone, Europol estimates such scams cost victims more than €200 million annually. These networks build convincing websites, stage professional interviews, and send official-looking contracts. Applicants are asked for their CV, home address, and then, at the end, comes a request for a small "processing fee" or "salary deposit." But this payment is irrelevant. The real prize is what they've already taken: a complete identity profile, neatly packaged and ready to sell.

These aren't lone scammers hunting for quick cash. They are organized networks running industrial-scale data operations. And once your information is in their hands, General Data Protection Regulation (GDPR)'s much-celebrated "right to erase" offers little comfort. How do you delete data that's already been copied, resold, and stored on servers in multiple countries? You cannot. It is gone, yet still alive, circulating in a black market you will never see. Emerging threats like these exploit one of the most vulnerable groups in our digital economy: young people entering the workforce for the first time, often unaware that the biggest risk they face is not job application rejection, but identity theft.

Until that moment, I had only a vague idea of how personal and irreversible data theft could be – but I learned almost overnight just how real it was

The danger is not just that people are tricked, it is that they're gradually conditioned to trust processes that quietly harvest their personal information. Job applications are one example, but similar dynamics emerge when interacting with AI chatbots. Scientific American reported that 72% of teenagers use chatbots like ChatGPT, designed to provide human-like interaction. They respond instantly, without judgment, and are available at any hour. Many teenagers online describe them as "parental figures," always there to listen. In some cases, chatbots end up knowing more about a person's fears, relationships, and secrets than the people closest to them. That trust might feel harmless, but AI is purely a technological tool with no human feelings and no enforceable duty of care. Conversations can be stored, retrieved, and sometimes made accessible to authorities or third parties. Unlike humans, AI cannot guarantee confidentiality. A Stanford study that same year found that over 60% of young users had disclosed sensitive personal information to a chatbot, often more than they had shared with any human. For some, the chatbot holds more truth about their lives than anyone they know, and that truth sits unprotected on a server they will never see.

This vulnerability is predictable when responsive, human-like tools are placed in the hands of isolated individuals. I, too, rely on AI as a law student, it is an incredibly efficient tool, but unlike human interactions, there is no legal guarantee or universally enforceable ethical framework ensuring confidentiality. Sensitive disclosures about abuse, mental health, sexuality, or political views can be recorded, analyzed, and subpoenaed. Technologies designed to feel trustworthy can quietly exploit that trust if safeguards are not clear, even as they provide real utility.

While these digital risks play out in the intimate spaces of our screens, the impact of AI and

digital technologies extends beyond our screens into the physical world. Every digital interaction runs on infrastructure consuming vast amounts of electricity and water; a single hyperscale data center can use more than 1.5 million gallons daily. Facilities are also often located near communities with limited economic or political influence, placing environmental and economic burdens on residents who have little say in these decisions.

This is not a problem unique to AI – digital infrastructure more broadly carries real-world costs, frequently borne by those least able to benefit from the technologies they support.

This raises urgent environmental justice questions: should the benefits of AI innovation come at the expense of communities that shoulder its environmental costs but reap few of its rewards? When innovation draws its power from the very communities it leaves behind, it no longer represents progress and starts becoming another form of exploitation.

Some defend this as a necessary cost, the energy required by technologies that could solve problems far bigger than itself. But that calculation rarely accounts for the human geography of its impact. The benefits are global, the burdens are local. And when infrastructure choices deepen inequities, this becomes an ethics concern, because the systems we build to protect and connect people should not harm the very communities they rely on.

These are not separate problems. The thread running through these examples is that they emerge quietly, in forms that seem innocuous. Trust is exploited, anonymity eroded, and costs, whether in stolen identities, unprotected personal confessions, or environmental strain — are displaced onto communities, including those with limited economic or political leverage. If cybersecurity's goal is not only to defend systems but to protect people, then policymakers and industry leaders must treat these subtle threats with the seriousness they deserve. We are entering an age where the most dangerous cyber threats may not be the loudest, but the ones that blend in so well that by the time we realise what is happening, they have already taken root.

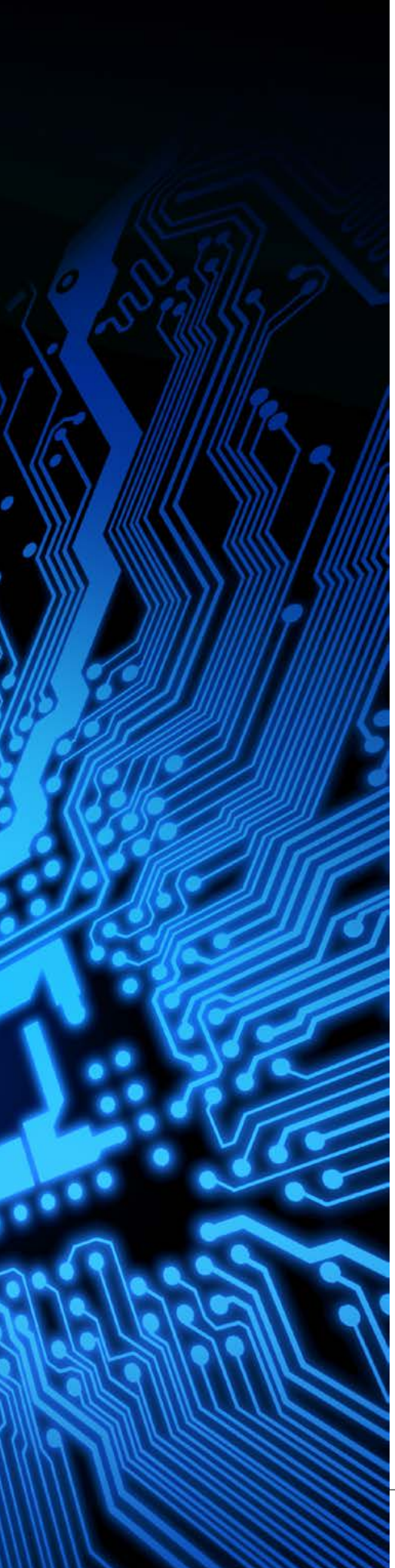
"

INTERPOL's 2023 cybercrime report found that recruitment scams targeting young job seekers increased by more than 30% in a single year. In the EU alone, Europol estimates such scams cost victims more than €200 million annually

ABOUT THE AUTHOR

Phoenix Omwando is a 20-year-old law student in the Netherlands, specializing in international and human rights law. Her work combines personal experience with cyber exploitation and academic research, focusing on emerging digital threats and environmental justice. She is driven by the belief that technology should protect, not exploit, and that innovation must never come at the expense of vulnerable communities. Phoenix seeks to bridge the gap between policy and lived reality, ensuring that those most affected by global decisions are heard, from local advocacy forums to the international stage.

Firewalls and antivirus software have become the Maginot Line of the digital age – static, brittle and ill-suited bulwarks against sophisticated threats.



Europe's invisible battleground: safeguarding the commons in an era of borderless threats

by Aiko Yeo

When ransomware crippled Ireland's national health service in 2021, the nation's lifelines buckled. Monitors went dark, ambulances were diverted, and vital medical records vanished behind encrypted walls. A subsequent government inquiry delivered a brutal verdict: firewalls and antivirus software have become the Maginot Line of the digital age – static, brittle and ill-suited bulwarks against sophisticated threats.

In Western Europe, such episodes are no longer outliers. They are a lived reality. The 2025 Iberian Peninsula blackout illustrates this point: a single breach can cascade across borders and morph into a continental crisis. Herein lies a paradox: digital systems, conceived as the backbone of rights and public utility, can just as quickly become a conduit for their erosion.

When a corrupted electoral register denies citizens the vote or a hacked asylum database strands refugees in legal limbo, the damage is not merely operational. These seismic shocks strike at the core of the region's total defence, where sovereignty, commerce and human rights are woven into the same fabric to secure society.

Against this backdrop, both Brussels and London have advanced flagship regulatory regimes aimed at fortifying Europe's digital ecosystem and softening the blow of the next disruption. This article first charts the protections within each framework and finds that

distinct legal dialects are shaped by policy priorities. Yet, despite divergent legal trajectories, a closer investigation reveals a shared two-pronged ambition. First, to make systemic cyber resilience a default design principle, rooted in resilience engineering and critical infrastructure protection theory. Second, to assign the heaviest duties to the most consequential actors, guided by proportionality, tempered by the precautionary principle, and anchored in a robust framework of duty of care.

Drawing the digital perimeter

Every credible cybersecurity framework first asks a deceptively simple question: Who and what merits protection? The answer reveals far more than technical scope; it exposes a government's priorities and its chosen balance between economic freedom, state oversight and individual rights.

In Brussels, that answer is unapologetically broad. *The Network and Information Security Directive (NIS2) Directive*[1] sweeps "essential" and "important" entities across 18 sectors into its orbit — not just power grids and telecoms, but public administrations, hospitals, managed service providers, and data centres. It signals a deliberate policy shift: resilience is framed not as a perk for critical industries but as a public good embedded in Europe's rights architecture. By incorporating these entities into the sphere of regional cyber defence, Brussels affirms that uninterrupted access is as much about equality as it is about economic necessity. A breach anywhere in any of these nodes then demands the same urgency typically reserved for a public health emergency, or a food safety crises.

Herein lies a paradox: digital systems, conceived as the backbone of rights and public utility, can just as quickly become a conduit for their erosion

The accompanying *Cyber Resilience Act*[2] pushes the perimeter outwards by setting horizontal security standards for any product with "digital elements". Its strategy is to confront systemic risk at the source, shifting accountability upstream to those best positioned to mitigate it, namely manufacturers and suppliers. Supplier-risk management is recast from a peripheral corporate afterthought into a binding legal mandate, harmonising industry practice with the EU's normative commitment to public protection.

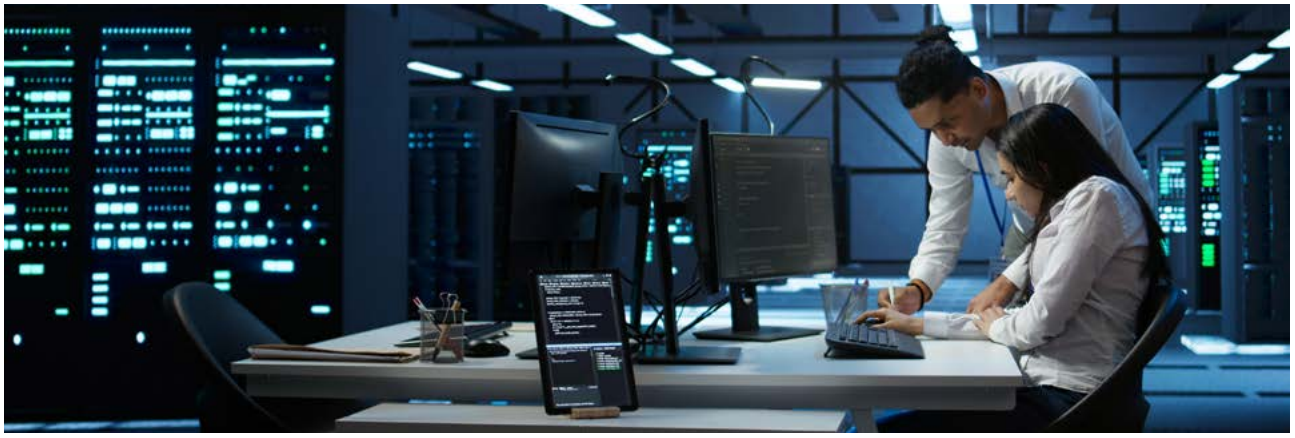
Meanwhile, London first borrowed Brussels' legal blueprint, then reshaped it for a post-Brexit regulatory landscape. The *NIS Regulations*[3] initially targeted Operators of Essential Services and a small cadre of digital service providers. That foundation will be reinforced by the forthcoming *Cyber Security and Resilience Bill,*[4] which widens the net to capture "designated critical suppliers": a sweeping category that includes large cloud operators and data-centre owners. In today's digital economy, these corporate gatekeepers control chokepoints as strategically vital as any power grids or

1 Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive) [2022] OJ L333/80.

2 Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act).

3 Network and Information Systems Regulations 2018, SI 2018/506.

4 Cyber Security and Resilience (Network and Information Systems) Bill (HC Bill 329, 2024-26).



water networks. From a rights perspective, their stability is not a parallel concern to public infrastructure, but inseparable from the economic and civic life they underpin.

Different scripts, same play

Europe's cybersecurity policy may speak in different dialects, yet each is a defence tactfully crafted to weather the same storm. Doctrinally, Brussels's model rests on the principle of universalised resilience, treating cybersecurity as a derivative of fundamental rights and a pillar of market integrity. The *NIS2 Directive* legislates for breadth, embedding cybersecurity into the single market alongside privacy and environmental safeguards. The *Cyber Resilience Act* pushes further, setting horizontal security obligations for corporate and public operators alike under a unified defensive perimeter. Thick regulation is the norm, with harmonised, binding duties rooted in a supranational commitment to equal protection.

> **Brussels's model rests on the principle of universalised resilience, treating cybersecurity as a derivative of fundamental rights and a pillar of market integrity**

In contrast, London's post-Brexit approach prizes proportionality and targeted statutory reach. The forthcoming *Cyber Security Resilience Bill* marks a deliberate recalibration, homing in on critical fault lines that could trigger systemic disruption. This thin constitutionalism values adaptability and sectoral tailoring over, seeking just enough cross-border coherence to stay aligned with Europe's defensive baseline.

Digging deeper, the divergence is more philosophical than operational. Brussels frames resilience as the latest pillar of continental rights; London grounds its scope in the twin imperatives of public stability and economic security. Nonetheless, both rest on the same normative conviction: resilience as a constitutional obligation, not a mere technical benchmark. In both frameworks, the heaviest duties fall on the most pivotal actors whose failures could imperil the commons, because defending against amorphous threats demands the strongest shoulders.

Defending the digital commons

In Europe's tangle of cables, clouds and codes, national boundaries become a relic of defences. Brussels and London may speak in different legal vernacular, but they share the same baseline: resilience as the norm, not the exception. An underlying impulse further binds both frameworks: to

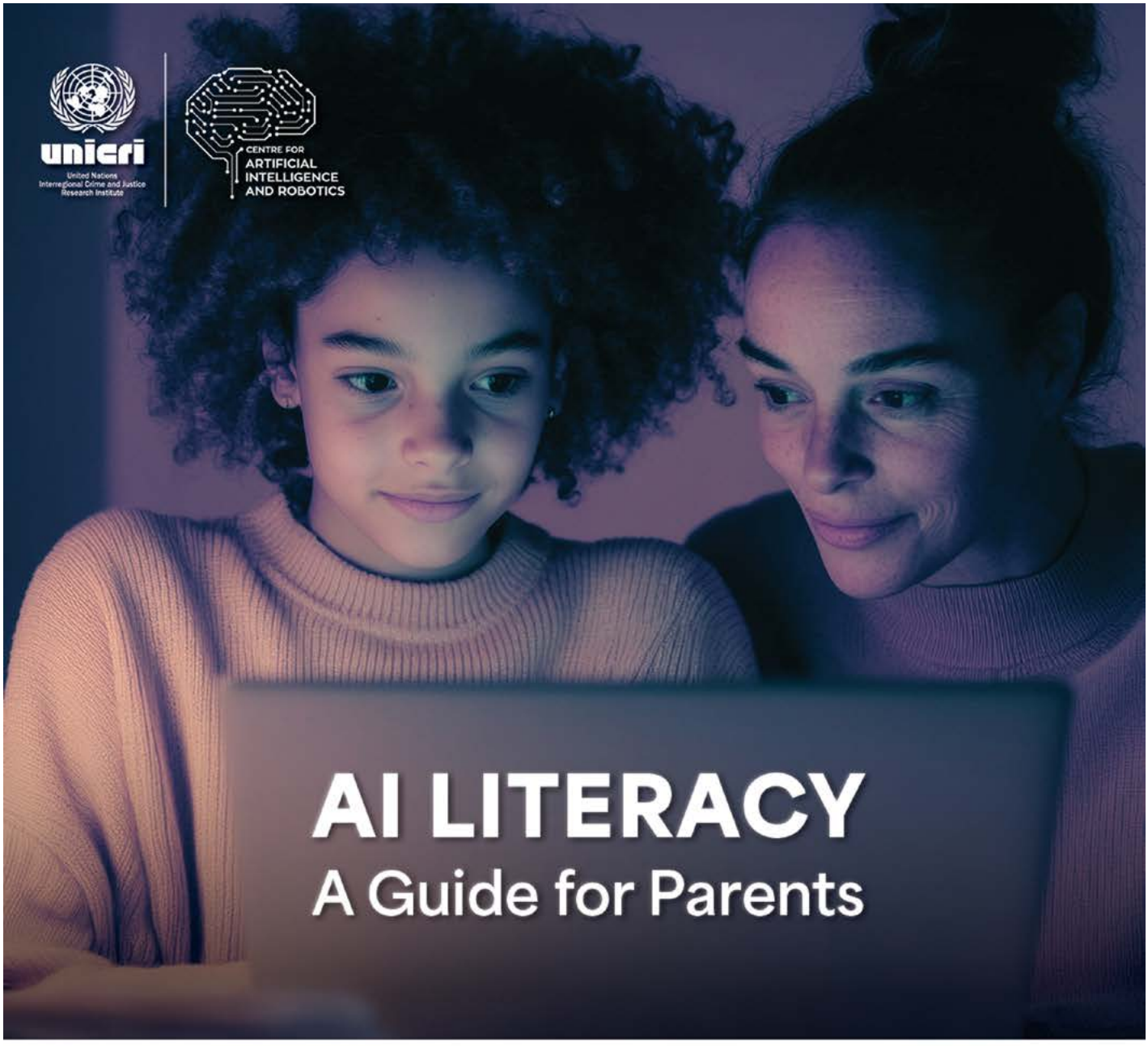
embed security deep within the arteries of public life, and to hold the most powerful actors to the heaviest duties of care.

But laws alone are fortifications without sentries: imposing on paper, porous in practice. When push comes to shove, neither Brussels' broad canopy nor London's tailored scaffolding will hold without the mortar of trust, in two senses of the word. First, interstate trust to trade intelligence when ramparts are tested, while resisting the reflex of political pride. Second, public trust in institutions to respond decisively, transparently and without deflection when cracks appear. In cybersecurity, as in diplomacy, the surest defence is not the walls nor the machinery, but the trust that stabilises it against fracture under the next shock.

Policymakers should resist treating Brussels' universalised resilience and blanket coverage as a separate track from London's targeted agility. Read together, both approaches offer complementary strengths: Brussels embeds security as a universal right, while London sharpens resources on the most consequential vulnerability. Yet, a hybrid system must confront the accompanying practical fissures: fragmented CERT-to-CERT communication, inconsistent reporting thresholds, intelligence silos that delay escalation, and the bureaucratic drag that slows real-time information sharing. A more effective course would be to shape a model that taps into the merits of both – common baseline standards, integrated supply-chain oversight, continuous joint exercises, and real-time intelligence exchanges – to stitch these defences into a system that can anticipate shocks. In the age of borderless threats, the only wrong move is to defend alone.

ABOUT THE AUTHOR

Aiko Yeo is a penultimate year LLB Law student from the London School of Economics and Political Science. As an incoming legally-trained policy analyst under Singapore's Ministry of Defence, she is committed to developing a nuanced understanding of the nexus between law, geopolitics and public policy. Aiko's academic and professional pursuits are united by a dedication to safeguarding national interests while strengthening the legal frameworks underpinning regional and global stability.


unicri
United Nations
Interregional Crime and Justice
Research Institute
CENTRE FOR
ARTIFICIAL
INTELLIGENCE
AND ROBOTICS
AI LITERACY
A Guide for Parents


unicri
United Nations
Interregional Crime and Justice
Research Institute
CENTRE FOR
ARTIFICIAL
INTELLIGENCE
AND ROBOTICS
RESPONSIBLE
AI INNOVATION IN
LAW ENFORCEMENT
AI Toolkit
Not Just Another Tool
Demographic profile
of the respondents
Funded by
the European Union

United Nations
Interregional Crime and Justice
Research Institute

GHENT
UNIVERSITY