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