BSAMUN 2025

Preventing Exploitation of Generative AI by Hostile Parties

General Assembly 1

President Chair:

Deputy Chair:

INTRODUCTION

Generative AI is moving at an unprecedented rate and we're in a new era of technological possibilities and human productivity. Generative AI refers to algorithms that can produce realistic text, images, audio, video, etc, on their own and is being used in creative industries and medical research. But this powerful technology is also a huge risk when used by bad actors for malicious purposes like disinformation, deep fakes or cyber attacks.

General Assembly Committee I (GA1) is tasked with stopping the misuse of generative AI while preserving its benefits. As these systems get more advanced the world is faced with a tightrope walk: harness the power of generative AI while mitigating the risks. This report looks at the core issues around preventing generative AI exploitation, the intersection of technological innovation, national security, ethics and international cooperation.

Through this research report, we aim to show nuanced insights that will enable the GA1 to develop effective policies and frameworks to address this pressing global security challenge. The report will analyze risks, evaluate existing international efforts, and propose strategies to ensure generative AI remains a force for progress rather than harm.

KEY TERMS

Artificial intelligence (AI): A collection of technologies that provide computers the capacity to carry out a wide range of sophisticated tasks, such as seeing, comprehending, and translating written and spoken language, analysing data, formulating suggestions, and more.

Generative AI: A form of AI that can produce original ideas and content, such as dialogue, stories, films, images, and music. It is capable of learning any complex subject, including biology, chemistry, programming languages, art, and human language.

Adversarial AI: A component of machine learning in which malicious parties consciously try to interfere with AI systems' ability to function. The purpose of adversarial attacks is to trick AI systems into making erroneous or unexpected predictions or judgements.

Safety By Design: A proactive approach to embedding safety measures into AI systems from the outset of their development

AI-generated child sexual abuse material (AIG-CSAM): Visual depictions of sexually explicit conduct involving minors created using AI technologies

Deepfakes: pictures, videos, or audio that have been altered or produced with the use of artificial intelligence algorithms and may show actual or fictional individuals.

AI Ethics: Discusses the moral implications of AI development and use, such as concerns about prejudice, privacy, and accountability.

AI Governance: The collection of procedures, guidelines, and resources that unite various stakeholders from data science, engineering, compliance, legal, and business teams to guarantee that AI systems are developed, implemented, and operated in a way that optimises benefits and minimises risks.

GENERAL OVERVIEW

The hostile use of generative AI has serious global implications for security, information integrity, democratic processes, and human rights. Misuse opportunities have grown exponentially along with the development and spread of these technologies, raising significant challenges in various domains.

Generative AI can create a huge amount of believable false information in order to manipulate the opinion of large groups of individuals and weaken trust in institutions. Thus, the ability to create realistic images and videos at scale drives down the line between what is real and what is not. Active use of this type of AI by hostile actors will enable highly personalized and convincing phishing emails, social media posts, or voice recording manipulation, which manipulates individuals and organizations.

One of the most concerning examples of generative AI security risks is the proliferation of AI-generated child sexual abuse material (AIG-CSAM). The Internet Watch Foundation (IWF) reported a 6% increase in the number of reports confirmed as containing criminal AI child sexual abuse material in the past six months since Octobert 2024 compared to the previous 12 months. This shows a growing threat that underscores the urgent need for proactive measures by the GA1 to combat the misuse of AI technologies.

To address these challenges, the concept of "Safety by Design" has been created as a proactive approach to embedding safety measures into AI systems from the outset of their development. This strategy aims to mitigate risks associated with generative AI by incorporating safeguards and ethical considerations throughout the design and implementation process.

Content provenance solutions have become vital differentiators between what is real and what is not from AI-generated content. New provenance methods, including tamper-resistant watermarking and detection classifiers, are being developed by organizations such as OpenAI in order to enhance the integrity of digital content. These are tools that will also be more resistant to the removal of signals related to the origin of content, helping to combat the spread of misinformation and deep fakes.

The robust testing methodology, including red-teaming and stress testing, can help identify the potential biases and vulnerabilities of AI systems. These approaches involve testing of AI models to understand their capabilities, particularly in producing abusive or harmful content.

Besides that, there are serious concerns regarding the security of the AI systems themselves, including model theft and reverse engineering. Most recently, ways to steal models without hacking were

demonstrated by researchers, showing that more protection needs to be given to AI model intellectual properties and sensitive data.

Against these challenges, the continuous efforts of technology companies and organizations include collaborations with anti-abuse organizations and the development of industry-wide standards. For example, OpenAI has joined the Coalition for Content Provenance and Authenticity (C2PA) to contribute to the development of open technical standards for certifying the source and history of media content.

With these challenges, the global community is faced with the growing requirement of an approach that is multi-layered in technology solutions, policy frameworks, and international cooperation as a sine qua non to prevent hostile parties from exploiting generative AI.

MAJOR PARTIES/COUNTRIES INVOLVED

United States: The U.S. remains a dominant force in the generative AI sector, accounting for the majority of global investment in this area. American companies continued to lead in generative AI funding in 2024, taking the lion's share of this sector.

China: While enacting laws to control their usage, China is avidly seeking developments in generative AI technologies. The Chinese government is a major player in the global AI arena because it is committed to making sure that advancements in AI are in line with societal stability and national goals. China is also advancing its AI capabilities through state-directed initiatives as part of its 14th Five-Year Plan.

The European Union: Through the EU AI Act, which creates extensive rules for AI development and use, the European Union plays a critical role in preventing hostile actors from exploiting generative AI. The EU seeks to protect citizens from possible misuses of generative AI by enforcing stringent regulations on high-risk applications and encouraging openness and responsibility.

The United Kingdom: In 2024, the UK is enhancing its AI capabilities through strategic investments and the introduction of the AI Bill, which focuses on regulating advanced generative AI models. The government has committed over £100 million to support AI innovation and is establishing new initiatives, such as the AI and Digital Hub, to foster a favorable environment for development.

TIMELINE OF KEY EVENTS

1949: Alan Turing introduces the Turing Test which was a test of a machine's ability to exhibit intelligent behaviour equivalent to that of a human. The event established foundational concepts in artificial intelligence that influence discussions about AI's capabilities and ethical considerations.

1956: The Dartmouth Conference is held, marking the birth of AI as a field of study and catalyzing research and development in AI technologies.

1997: IBM's Deep Blue defeats world chess champion Garry Kasparov, showcasing AI's potential and raising public awareness about the implications of advanced AI systems.

2015: Facebook introduces and rolls out DeepFace, a facial recognition system that utilizes deep learning techniques to identify human faces in digital images, achieving an accuracy rate of approximately 97.35%, which is comparable to human recognition capabilities. The system raises concerns about privacy and potential misuse of AI technologies.

2022: OpenAI introduces ChatGPT, which becomes well-known very fast. This raises questions regarding the moral ramifications of generative AI, particularly its potential for abuse in disinformation efforts. Since students may use ChatGPT to complete assignments dishonestly, compromising the learning process, educators and legislators are concerned about academic integrity due to the tool's fast popularity.

March 1, 2024: The European Parliament approves the EU AI Ac - with a vote of 523 for, 46 against, and 49 abstaining- , which aims to regulate AI technologies through a risk-based approach to prevent exploitation by hostile parties. This legislative move is seen as a proactive measure to establish a framework that balances innovation with safety, ensuring that AI developments align with fundamental rights and societal values.

July 12, 2024: The EU AI Act is published in the Official Journal of the European Union, marking the start of its implementation timeline. This publication lays the groundwork for member states and stakeholders to get ready for the upcoming legislation

August 1, 2024: The EU AI Act officially enters into force, establishing a regulatory framework for AI that includes provisions to prevent misuse. As a result of this enforcement, businesses and organisations must now evaluate the risk levels of their AI systems and follow rules intended to reduce any potential negative effects of their technologies.

June 10, 2024: A report by the International Centre for Counter-Terrorism (ICCT) highlights the need for collaboration among policymakers, law enforcement, and civil society to develop robust strategies to counter the misuse of generative AI by terrorist groups. The report emphasizes enhancing regulatory frameworks and investing in AI detection tools to mitigate risks associated with generative AI exploitation.

October 28, 2024: Forecasts point to an increase in cybercrime using generative AI, especially in social media settings. It is anticipated that criminals would employ AI for focused impersonation attacks, which will raise the complexity and scope of fraud and scams and increase the demand for efficient defences.

Current: With more and more accounts of cybercriminals using these technologies, it is clear from ongoing conversations that generative AI is being misused. The use of generative AI to produce deep fakes, complex phishing techniques, and automated cyberattacks are noteworthy issues. To reduce the dangers of unwanted data sharing and possible exploitation by adversaries, organisations are advised to strengthen data security protocols and implement more stringent controls over the use of AI tools.

UN INVOLVEMENT AND RELEVANT RESOLUTIONS

The Interim Report: Governing AI for Humanity (December, 2023): The report addresses the urgent need for comprehensive governance frameworks to manage the risks associated with artificial intelligence, including its potential misuse by hostile parties. It emphasizes global cooperation, ethical principles, and the establishment of norms to ensure that AI technologies benefit all of humanity while safeguarding human rights.

Security Council Meeting on AI (December 19, 2023): In its first formal meeting on artificial intelligence, the UN Security Council discussed the dual risks and rewards of AI technologies.In order to stop the exploitation of AI to spread hate speech, prejudice, and totalitarian monitoring, Secretary-General António Guterres emphasised the critical need for ethical governance frameworks. The UN's dedication to tackling the issues raised by generative AI on a global scale was emphasised at this meeting.

UN General Assembly Resolution on AI (March 21, 2024): The UN approved its first significant resolution on artificial intelligence, urging member states to develop "safe, secure, and trustworthy AI systems" while respecting human rights and international law and “to govern this technology rather than let it govern us”. This resolution emphasizes the importance of ensuring that AI technologies do not pose undue risks to human rights, particularly for vulnerable populations. It calls for states to refrain from using AI systems that cannot comply with international human rights law or that present significant risks to rights enjoyment.

High-Level Advisory Body on Artificial Intelligence Report (September 2024): The United Nations Secretary-General’s High-level Advisory Body on AI’s Final Report, “Governing AI for Humanity,” was released in September 2024 and builds on months of work, including extensive global consultations, and

the publication of an interim report in December 2023.

PREVIOUS ATTEMPTS TO SOLVE THE ISSUE

Ethical AI Principles (2020): The CEB endorsed the Principles for the Ethical Use of Artificial Intelligence in the United Nations System.

UNESCO AI Ethics Recommendation (November 2021): UNESCO's General Conference adopted the Recommendation on the Ethics of Artificial Intelligence, which was later translated into principles for UN system entities .

AI Capacity Development (April 2019): The CEB endorsed a system-wide strategic approach and road map for supporting AI capacity development, focusing on developing countries and ethical AI application

HLCM Task Force on AI (October 2023): Established to develop a system-wide framework for AI use in the UN and promote mechanisms for pooling technical capacity and knowledge sharing on AI .

OECD and UN Collaboration (September 2024): The two organizations announced enhanced collaboration on global AI governance, focusing on regular science and evidence-based AI risk and opportunity assessments.

POSSIBLE SOLUTIONS

Stopping hostile actors misusing generative AI requires a multi-layered approach that balances innovation with security. A first step is to establish international standards for the ethical development and use of generative AI. In GA1 member states can agree on guidelines that require transparency in AI systems, so developers have to label synthetic content and embed filters in content. This would make it harder for AI to be used for malicious purposes like disinformation campaigns or deepfakes. And these standards can also integrate ethical considerations into the development process.

Public education and awareness is key to stopping the misuse of generative AI. Member States should invest in programs to increase AI literacy so individuals can critically assess and identify AI generated content. Including AI in national school curricula and targeted training for public officials, media professionals and educators can make societies more resilient to disinformation and manipulation. And implementing robust verification systems is key to counter the spread of malicious AI generated media. Using technology like blockchain to authenticate digital content can stop the spread of harmful synthetic media and hold creators accountable.

International cooperation is key to stopping the misuse of generative AI in cyber attacks. A UN led

framework for sharing intelligence on emerging AI threats can help member states stay ahead of hostile actors. This should be complemented by capacity building programs especially in developing countries to strengthen global cybersecurity infrastructures. Member States should also promote accountability in the private sector by incentivizing ethical AI development and requiring transparency in vulnerability reporting.

Finally, stopping the misuse of generative AI requires a robust enforcement mechanism. The General Assembly could create a framework for investigating and sanctioning individuals, organizations or states found to be using AI for malicious purposes. To coordinate this a specialized UN oversight body like a Global AI Governance Forum could be established. This would be a central authority to monitor AI applications, mediate disputes and facilitate international cooperation. By doing so the global community can reap the benefits of generative AI while preventing its misuse and ensuring it aligns with our shared values of peace and security.

BIBLIOGRAPHY

"AI and Digital Hub." *Www.drcf.org.uk*, 11 Sept. 2024, www.drcf.org.uk/ai-and-digital-hub/.

ARPU. "Generative AI Funding Soars to Record $56 Billion in 2024." *ARPU*, Jan. 2025, https://doi.org/10879063107/bkUqCNzWx8UDEMOoxcMo.

"Artificial Intelligence | United Nations - CEB." *Unsceb.org*, unsceb.org/topics/artificial-intelligence.

Artificial Intelligence: High-level Briefing. "Artificial Intelligence: High-Level Briefing." *Security Council Report*, 2024, www.securitycouncilreport.org/whatsinblue/2024/12/artificial-intelligence-high-level-briefing.pp

AWS. “What Is Generative AI? - Generative Artificial Intelligence Explained - AWS.” *Amazon Web Services, Inc.*, 2024, aws.amazon.com/what-is/generative-ai/.

“Credo AI - What Is AI Governance?” *Credo.ai*, 2024, www.credo.ai/glossary/ai-governance. 8

Dredge, Stuart. “Global Generative-AI Funding Nearly Tripled in 2024 to $56bn.” *Music Ally*, 6 Jan. 2025, musically.com/2025/01/06/global-generative-ai-funding-nearly-tripled-in-2024-to-56bn/.

European Parliament. “EU AI Act: First Regulation on Artificial Intelligence.” *European Parliament*, 18 June 2024, www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation on-artificial-intelligence.

Google Cloud. “What Is Artificial Intelligence (AI)?” *Google Cloud*, 2023, cloud.google.com/learn/what-is-artificial-intelligence.

“How ‘Safety by Design’ Principles Aim to Change the AI Industry | CameraForensics.” *Cameraforensics.com*, 2024, www.cameraforensics.com/blog/2024/05/23/how-safety-by-design-principles-aim-to-cha nge-the-ai-industry/.

Nations, United. “AI Advisory Body.” *United Nations*, Dec. 2023,

www.un.org/en/ai-advisory-body.

"OECD and UN Announce next Steps in Collaboration on Artificial Intelligence." *OECD*, 2024, www.oecd.org/en/about/news/press-releases/2024/09/oecd-and-un-announce-next-steps-in-collaboration-on-artificial-intelligence.html.

Office, Intellectual Property. "Copyright and Artificial Intelligence." *GOV.UK*, 17 Dec. 2024, www.gov.uk/government/consultations/copyright-and-artificial-intelligence.

---. "Copyright and Artificial Intelligence." *GOV.UK*, 17 Dec. 2024, www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence.

United Nations. "General Assembly Adopts Landmark Resolution on Artificial Intelligence | UN News." *News.un.org*, 21 Mar. 2024, news.un.org/en/story/2024/03/1147831.

"United Nations AI Resolution: A Significant Global Policy Effort to Harness the Technology for Sustainable Development | IHEID FC." *Executive.graduateinstitute.ch*, 6 May 2024, executive.graduateinstitute.ch/communications/news/united-nations-ai-resolution-significant-global-policy-effort-harness.

"United Nations Security Council Arria-Formula Meeting on Artificial Intelligence | UN Web TV." *Webtv.un.org*, 19 Dec. 2023, webtv.un.org/en/asset/k1g/k1glfm55l2.

"What Is Adversarial AI in Machine Learning?" *Palo Alto Networks*, www.paloaltonetworks.com/cyberpedia/what-are-adversarial-attacks-on-AI-Machine-Learning.

White & Case. "AI Watch: Global Regulatory Tracker - United Kingdom | White & Case LLP." *Www.whitecase.com*, 13 May 2024, www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-kingdom.

Wiggers, Kyle. "Generative AI Funding Reached New Heights in 2024 | TechCrunch." *TechCrunch*, 3 Jan. 2025, techcrunch.com/2025/01/03/generative-ai-funding-reached-new-heights-in-2024/.

1
1