General Assembly

AI-driven Disinformation and Deepfake Warfare: A global threat to privacy, peace and political integrity.

1. *History of the committee*

The General Assembly was established when the UN Charter entered into force on 24 October 1945 and held its first session on 10 January 1946 at Methodist Central Hall in London. From the outset, it served as the UN's main deliberative, policymaking, and representative organ, giving equal voice to all Member States.

As the Assembly's central functions evolved, it took charge of the UN budget, appointed non-permanent Security Council members and the Secretary-General, reviewed reports from UN bodies, and created subsidiary organs to tackle new global challenges. It meets each year in New York from September through January and can reconvene in special or emergency sessions to address urgent crises.

Major achievements of the General Assembly include the adoption of the Universal Declaration of Human Rights in 1948, the creation of UNICEF in 1946 and UNDP in 1965 to institutionalize humanitarian and development work, and the 2015 unanimous agreement on the 2030 Agenda for Sustainable Development with its 17 Sustainable Development Goals.

2. *Introduction*

AI-driven disinformation and deepfake warfare harness cutting-edge machine learning to fabricate hyperrealistic videos, audio, and imagery that impersonate real individuals and events. These synthetic media can be produced and distributed at scale, making it increasingly difficult for citizens, media outlets, and policymakers to distinguish fact from manipulation.

The rapid spread of deepfakes poses a grave threat to personal privacy, democratic processes, and international peace. By eroding trust in official communications and inflaming social tensions, deepfake campaigns can spark violence, interfere in elections, and undermine the integrity of governments. As the world's principal deliberative body, the General Assembly must guide member states in establishing norms, legal frameworks, and cooperative safeguards to detect, deter, and penalize malicious uses of AI-generated content.

Beyond content creation, AI-driven platforms amplify disinformation through hyper-personalized targeting and algorithmic recommendation systems. By segmenting audiences based on their digital footprints, these systems spread tailored falsehoods into echo chambers, eroding shared facts and fueling polarization

The right to privacy has also come under siege. AI-enabled deepfake pornography and identity theft exploit intimate data without consent, leaving victims vulnerable to blackmail, reputational harm, and emotional distress. Such violations challenge existing legal frameworks and demand urgent remediation.

Political integrity and peace are similarly jeopardized. States and non-state actors have weaponized AI-driven disinformation to interfere in elections, undermine public trust in democratic institutions, and stoke social unrest. While memes and basic "fake news" have already influenced major votes, deepfakes pose an even greater risk of decisively swaying public opinion through fabricated events that never occurred.

---

1. *Historical context*

Long before the internet, authorities and interest groups relied on print media, radio broadcasts, and psychological operations to influence public opinion and discredit opponents. These analog methods laid the groundwork for today's sophisticated information campaigns, demonstrating the enduring power of narrative control in shaping perceptions.

The emergence of social media in the early 2000s transformed disinformation from isolated smear tactics into a global phenomenon. Platforms such as Facebook and Twitter enabled rapid dissemination of memes, fabricated news articles, and targeted messages. Notably, Russia's meme-based interference in the 2016 U.S. presidential election and China's coordinated online narratives during the Hong Kong protests illustrated how digital networks could be weaponized to polarize societies and erode trust in institutions.

Around 2014, breakthroughs in machine learning introduced generative adversarial networks (GANs), which provided the technical basis for creating deepfakes—synthetically altered videos or audio clips that convincingly mimic real individuals. By 2017, hobbyist communities on forums like Reddit had coined the term "deepfake" after sharing AI-generated celebrity videos, sparking widespread concern about the technology's potential misuse.

Over the past five years, user-friendly deepfake software and powerful large-language models have democratized synthetic-media creation. Mobile applications now enable real-time face swaps, while emerging text-to-video generators threaten to render entirely fabricated events indistinguishable from reality. This rapid proliferation raises the stakes, as anyone with a smartphone can potentially undermine privacy, destabilize communities, or manipulate political outcomes.

Understanding this progression from analog propaganda to AI-enabled disinformation is essential for designing effective countermeasures. As synthetic-media capabilities evolve, legal frameworks, technical detection tools, and comprehensive media-literacy initiatives must advance in tandem to safeguard individual privacy, preserve peace, and uphold the integrity of democratic processes.

2. *Current Issue*

The proliferation of AI-driven disinformation and deepfakes has turned once-obvious manipulations into seamless fabrications. Today's generative adversarial networks and

large language models can produce videos of public figures saying things they never uttered, or synthesize audio that perfectly mimics a real voice. Social media algorithms then propel this content into echo chambers, making falsehoods spread faster than fact checks can keep up. As a result, audiences face an erosion of shared reality, unsure which sources to trust.

This escalation poses an acute threat to both national security and individual privacy. Armed forces and intelligence agencies have integrated deepfake capabilities into their doctrines, viewing them as force multipliers in psychological operations and "false-flag" campaigns designed to justify aggression. Meanwhile, emerging research highlights how nonstate actors—from extremist groups to criminal gangs—leverage the same tools for blackmail, identity theft, and disinformation-driven extortion, imperiling the emotional well-being of private citizens. International bodies such as NATO and the United Nations University emphasize that without robust detection algorithms, stringent regulatory frameworks, and widespread media-literacy education, the very foundations of democratic discourse and public trust risk irreparable damage

State and non-state actors alike have incorporated deepfakes into their strategic playbooks. Major powers invest heavily in AI research to refine these tools: for example, Russia's 2025 federal AI strategy dedicated 7.7 billion rubles to bolster its propaganda and hybrid-warfare capabilities. Meanwhile, extremist groups and criminal networks exploit deepfakes for blackmail, identity theft, and targeted harassment. In regions with fragile institutions—such as parts of Sub-Saharan Africa—AI-driven falsehoods have already intensified political tensions and fueled local conflicts.

Efforts to combat this threat face three interrelated challenges. First, detection technologies struggle to keep pace with increasingly evasive deepfake algorithms. Second, legal and regulatory frameworks remain uneven: some nations impose transparency requirements on platforms, while others lack clear rules. Third, public awareness and media-literacy initiatives are not yet widespread enough to inoculate citizens against sophisticated forgeries. Addressing these gaps will require coordinated international action, combining robust technical defenses, harmonized regulations, and

sustained education campaigns to protect privacy, preserve social cohesion, and uphold the integrity of democratic processes.

3. *Past international actions*

Recognizing the profound risks deepfakes pose to democratic trust, the Brookings Institution in January 2023 recommended that the United States and its allies develop a formal code of conduct for government use of synthetic media. The proposed "Deepfakes Equities Process" would convene stakeholders across ministries, intelligence agencies, and civil-society representatives to weigh the strategic benefits of deploying deepfakes against the necessity of transparency and protection of civil liberties. Modeled on established cybersecurity equities processes, this approach seeks to ensure that any state-sponsored synthetic content adheres to international norms of responsible innovation and conflict de-escalation. By institutionalizing deliberation over deepfake use, democracies aim to lead by example, preserving a trustworthy information environment while retaining strategic flexibility.

On the regulatory front, the European Union's Digital Services Act obliges major online platforms to implement mechanisms for identifying, labeling, and tracking AI-generated content, alongside regular reporting of disinformation metrics. Several Member States have bolstered this framework by introducing criminal penalties for the malicious creation and dissemination of non-consensual synthetic media. Simultaneously, the Council of Europe is advancing amendments to its Convention on Cybercrime to explicitly cover deepfake-enabled offenses and to strengthen judicial cooperation against cross-border disinformation operations.

Multilateral bodies have also taken decisive steps. In 2021, UNESCO adopted the Recommendation on the Ethics of Artificial Intelligence, urging Member States to shield individuals from manipulative media, safeguard personal data, and invest in comprehensive media-literacy education. The 2023 G7 Hiroshima AI Process culminated in a declaration that emphasizes harmonized principles on AI transparency, accountability, and a shared resolve to counter the malicious use of synthetic media.

Together, these initiatives reflect a growing consensus that only coordinated, multi-stakeholder action can effectively defend privacy, uphold democratic trust, and sustain international peace.

4. Subtopics

- Proliferation and Non-Proliferation:
    - The diffusion of AI-driven disinformation tools and deepfake capabilities among state and non-state actors.
    - Understanding the vectors and incentives that accelerate spread, to inform targeted prevention strategies
    - Assessing how increased availability elevates the overall threat level and amplifies geopolitical risks.

- Technological Advancements and Ethical Concerns:
    - How breakthroughs in generative adversarial networks and large-language models have perfected synthetic media.
    - Evaluating the ethical implications of democratized deepfake creation, from non-consensual content to automated persuasion.
    - Identifying regulatory challenges posed by rapid innovation and setting standards for responsible AI development.

- Detection and Attribution Challenges:
    - The evolution of forensic and machine-learning methods to identify synthetic audio, video, and text.
    - Obstacles in tracing disinformation campaigns back to their originators, given anonymizing tools and proxy networks.
    - Building robust attribution frameworks that support legal accountability without compromising privacy rights.

5. *Positions*

Nations such as the United States, the European Union (led by Germany and France) and Canada share a commitment to uphold democratic values while countering malicious synthetic media. The US emphasizes a multi-stakeholder model, forging partnerships with tech companies and civil society to deploy detection tools and voluntary codes of conduct, while resisting binding treaties that could hinder innovation or free expression. Meanwhile, the EU is pushing for a legally binding UN framework mandating clear AI-content labeling, mandatory transparency from major platforms, and robust enforcement across member states. Canada, drawing on UNESCO's ethical AI principles, champions a "Digital Rights Charter" that fuses strict privacy safeguards with expedited legal remedies for victims of non-consensual deepfakes.

On the other end of the spectrum, Russia and China stress digital sovereignty and state-led governance of AI media. Russia rejects supranational oversight, insisting each government independently regulates deepfake threats through national security agencies, even as it faces accusations of weaponizing synthetic content. China similarly enforces stringent domestic guidelines for online platforms under a state-approved code of ethics, framing regulation as essential to social stability. Both Moscow and Beijing endorse UN-coordinated principles in theory, but reserve ultimate implementation and enforcement for their own authorities.

Emerging economies and regional coalitions call for capacity building, equitable tool access, and inclusive policymaking. India proposes non-binding international transparency guidelines paired with UN-backed media-literacy programs and a South–South technical assistance network. Brazil advocates a dedicated fund for developing states to build detection labs and train moderators through technology transfer. The African Union, led by Nigeria, demands UN resources for real-time threat monitoring and cyber-response centers. The Group of 77 & China coalition highlights the digital divide, calling for a UN special rapporteur on synthetic-media harms. South Korea offers to pilot an open-source deepfake-detection toolkit under UN supervision and urges the inclusion of a technical annex with standardized forensic protocols in any resolution.

6. *Guiding questions*

- What exactly do we mean by AI-driven disinformation and deepfake warfare?
- How should the committee define the scope of "synthetic media" versus traditional propaganda to ensure clarity in drafting resolutions?
- Which state and non-state actors are currently deploying deepfakes as tools of influence or coercion?
- What patterns have we seen in recent elections, conflicts, or social movements that can inform prevention and response strategies?
- What international legal principles and existing frameworks can be adapted or expanded to cover the malicious use of generative AI?
- Where are the gaps that a new convention or protocol must fill?

Bibliography

Shin, J. (2024). *AI and Misinformation | 2024 Dean's Report*. 2024.Jou.ufl.edu; University of Florida. https://2024.jou.ufl.edu/page/ai-and-misinformation

Sperling, J. (2024, August 5). *AI Misinformation: Challenges and Solutions for Businesses | Columbia Business School*. Columbia Business School. https://business.columbia.edu/insights/digital-future/ai-misinformation-challenges-and-solutions-businesses

Byman, D., Gao, C., Meserole, C., & Subrahmanian, V. S. (2023, January). *Deepfakes and international conflict*. Brookings. https://www.brookings.edu/articles/deepfakes-and-international-conflict/