Disarmament and International Security Commission (DISC)

Research Report

CACMUN

Forum: Disarmament and International Security Commission (DISC)

Issue: The Question of the Threats Posed by Weaponized Deepfakes in International Conflict

Student Officer(s): Jade Abouelkheir

Position: Deputy Chair

Introduction

Deepfakes emerged in the late 2010s as a form of synthetic media, leveraging deep learning to manipulate videos and audio in ways that are increasingly indistinguishable from real recordings. Originally used for entertainment, satire, and harmless novelty, deepfake technology quickly attracted global attention due to its potential for misuse. What started as altered celebrity videos on internet forums has evolved into a powerful tool capable of threatening truth, trust, and transparency in the international sphere. The exponential growth of generative AI has only accelerated this development, enabling the creation of highly realistic deepfakes at a fraction of the time and cost it once required. As a result, the manipulation of information is no longer confined to traditional propaganda tools but has entered a new, more dangerous era of digital deception.

The current landscape of international conflict is particularly vulnerable to the weaponization of deepfakes. Unlike conventional cyberattacks, deepfakes target perception, emotion, and credibility. A single convincing video can go viral in minutes, leading to mass panic, political unrest, or escalations in conflict. In warzones, deepfakes can be deployed to impersonate military leaders issuing false commands, claim fabricated atrocities by enemies, or forge diplomatic statements. These manipulations can distort the truth on the ground, hinder peace negotiations, or even provoke international retaliation. The absence of standardized global media verification systems and the fragmented digital governance landscape further intensify the challenge, making it difficult to detect and respond to deepfakes quickly and effectively.

Key stakeholders in this issue include technologically advanced nations such as the United States, China, and Russia, all of whom possess both the capabilities to create deepfakes and the incentives to protect themselves from them. International organizations such as the United Nations, NATO, and Interpol have acknowledged the implications of deepfakes for peace and security, but have yet to develop comprehensive multilateral responses. Private technology firms, particularly those operating social media platforms and AI development tools, also play a pivotal role, as they host, amplify, and moderate much of the content disseminated globally. Civil society, journalists, and at-risk populations are often the first affected by weaponized deepfakes and remain among the most important actors in building resilience through media literacy and advocacy. Tackling this issue requires urgent, coordinated efforts across borders and sectors to protect truth, promote trust, and uphold international stability.

Definition of Key Terms

Deepfake

A type of synthetic media created using artificial intelligence, particularly deep learning, to generate or manipulate audio, video, or images to convincingly misrepresent someone's likeness or voice.

Weaponization

The process of using a tool, technology, or idea as a means of harm, manipulation, or coercion, typically in a military, political, or psychological context.

Information Warfare

A conflict or campaign waged in the information environment, including the use of fake news, propaganda, cyber operations, and digital disinformation to influence, deceive, or destabilize.

Generative AI

Artificial intelligence systems, like GANs (Generative Adversarial Networks), capable of producing new content such as images, text, and video that mimic human-made media.

Misinformation vs. Disinformation

Misinformation is false information spread without intent to deceive; disinformation is false information deliberately spread to mislead or manipulate.

Digital Verification Infrastructure

Technologies and frameworks used to confirm the authenticity of digital content, including blockchain, digital watermarks, and fact-checking tools.

Background Information:

The origins of deepfake technology can be traced to academic research in machine learning, particularly the use of GANs developed around 2014. However, it'ss public emergence occurred in 2017 when Reddit users began generating convincing fake celebrity videos. As computing power became more accessible and open-source deepfake tools proliferated, the technology's use quickly expanded beyond novelty content into political and social spheres.

Concerns deepened in the early 2020s as deepfakes began appearing in election campaigns, social unrest, and conflict zones. In 2022, a deepfake video of Ukrainian President Volodymyr Zelensky allegedly telling troops to surrender circulated during the Russian invasion of Ukraine. Although quickly debunked, the incident illustrated the potential for such media to disrupt morale and confuse military operations. Other examples have included deepfaked audio of African politicians, falsified news broadcasts, and fake diplomatic statements posted on social media.

Governments, the private sector, and civil society have responded in various ways, but without a unified strategy. Some nations have passed laws criminalizing the malicious use of deepfakes, while others are developing detection tools using AI itself. Tech platforms like Meta and X (formerly Twitter) have introduced labelling and takedown policies, but enforcement remains inconsistent. Internationally, discussions about regulating synthetic media have taken place in forums such as the UN General Assembly and the G7, but no binding treaty or framework has emerged.

As conflicts become increasingly hybrid, combining kinetic, cyber, and informational tactics, the role of deepfakes is expanding. Their potential to manipulate populations, justify aggression, or provoke intervention is a growing concern for the international community.

Causes of the Issue:

The core causes of this issue lie in the rapid advancement of artificial intelligence technologies, the accessibility of open source deepfake tools, and the lack of international regulation or standardization for synthetic media. The digital arms race among major powers has incentivized the development of increasingly powerful AI capabilities, with few checks on dual-use technologies. Moreover, the widespread use of social media platforms allows deepfakes to spread uncontrollably before they can be verified or removed.

Geopolitically, deepfakes are attractive tools in asymmetric warfare and hybrid conflict, where states or non-state actors seek to destabilize opponents without direct military confrontation. The global lack of media literacy also contributes to the issue, as many populations remain vulnerable to manipulation and unable to distinguish between real and fake content.

Effects of the Issue:

- **On Individuals and Citizens:** Deepfakes have been used to harass activists, journalists, and politicians, undermine reputations, and incite violence. Misleading content can erode trust in institutions, stoke ethnic tensions, and trigger civil unrest.
- **On National Governments:** Deepfakes can discredit public officials, disrupt military communication, or interfere in democratic elections, weakening state legitimacy and public trust.
- **On International Peace and Security:** Weaponized deepfakes can exacerbate conflicts, spread propaganda across borders, and create confusion during crises. The threat of escalation from a falsified diplomatic or military statement poses serious risks to peace.

Major Countries and Organizations Involved

United States of America

As a global leader in both AI development and cybersecurity, the U.S.A is heavily invested in deepfake detection and countermeasures. It has passed limited legislation and has called for international cooperation but faces internal political divisions over tech regulation.

Russia

Russia has been implicated in disinformation campaigns using deepfakes and is believed to possess strong capabilities in digital psychological operations. It views information warfare as a core strategy in modern conflict.

China

China has implemented domestic laws to regulate deepfakes internally and has voiced concern about foreign misuse. However, its strategic use of synthetic media abroad remains opaque.

Europeanian Union (EU)

The EU has taken a leading role in proposing digital content regulation, including the Digital Services Act, which holds platforms accountable for disinformation. It promotes media literacy and transparency initiatives.

United Nations (UN)

The UN has warned about deepfakes in the context of peacekeeping, misinformation, and electoral interference. However, there is no binding international framework addressing their weaponization in conflict.

Timeline of Events

Date	Description of Event
June 10, 2014	GANs Developed: Ian Goodfellow and his colleagues introduce Generative Adversarial Networks (GANs), publishing the foundational paper that lays the groundwork for deepfake technology.
December 11, 2017	Reddit Deepfake Videos: A Reddit user named “deepfakes”

	begins posting AI-generated pornographic videos using celebrity faces, marking the first viral spread of deepfake content and prompting widespread media coverage.
March 16, 2022	Zelensky Deepfake Incident: A manipulated video of Ukrainian President Volodymyr Zelensky telling troops to surrender is uploaded and briefly aired on a hacked Ukrainian news website during the ongoing Russia-Ukraine war.
March 20, 2023	G7 Hiroshima Leaders Statement: G7 nations, meeting in Hiroshima, Japan, release a joint statement acknowledging the threat of AI-generated disinformation and deepfakes, calling for international frameworks on AI governance.
August 28, 2023	UNESCO Deepfake Guidelines Released: UNESCO publishes its "Guidelines for the Governance of Digital Platforms," outlining ethical standards and recommendations for countering deepfake technologies and disinformation.
February 6, 2024	U.S. National AI Deepfake Detection Challenge Launched: The U.S. Department of Homeland Security and DARPA launch a national competition to accelerate the development of AI tools capable of identifying and flagging deepfakes in real time.

Relevant UN Treaties and Events

UNESCO Ethical AI Guidelines (2023)

- **Full Name:** Recommendation on the Ethics of Artificial Intelligence
- **Date Adopted: November 23, 2021 (Implementation expanded in 2023)**
- **Summary:**
    - The first global standard-setting instrument on the ethical use of artificial intelligence, adopted by UNESCO and supported by 193 member states.
    - Outlines core principles such as transparency, accountability, human rights protection, and sustainability in AI development.

- Emphasizes the need to combat disinformation and ensure media authenticity, including through the verification of AI-generated content like deepfakes.
- Recommends national governments develop media literacy initiatives, legal safeguards, and detection tools.
- Implementation toolkits were launched in 2023 to guide states in aligning national policies with these ethical principles.
- Non-binding but widely recognized as a global benchmark for ethical AI governance.

UN General Assembly Resolution on Countering Disinformation (A/RES/76/227)

- **Full Name:** Countering Disinformation for the Promotion and Protection of Human Rights and Fundamental Freedoms
- **Date Adopted: December 24, 2021**
- **Summary:**
  - A non-binding resolution passed by consensus in the UN General Assembly.
  - Calls for international cooperation to counter the spread of false or manipulated information, especially through digital technologies.
  - Urges member states to promote access to reliable information, support fact-based journalism, and invest in digital literacy programs.
  - Acknowledges the threat of AI-generated disinformation, including deepfakes, to democratic processes and international peace.
  - Encourages states to develop legal and educational measures while respecting freedom of expression.
  - Frequently cited in global discussions on regulating harmful synthetic media.

International Covenant on Civil and Political Rights (ICCPR, 1966)

- **Full Name:** International Covenant on Civil and Political Rights
- **Date Adopted: December 16, 1966 (Entered into force: March 23, 1976)**

- **Summary:**
    - A legally binding multilateral treaty adopted by the UN General Assembly and ratified by over 170 countries.
    - Article 19 guarantees the right to freedom of expression and access to information.
    - Allows for restrictions only when necessary to protect national security, public order, public health, or the rights of others.
    - Cited in international debates on AI and media regulation, including the governance of deepfakes.
    - Requires that any restrictions on speech, including digital content moderation, be lawful, proportionate, and respectful of human rights.
    - Forms the legal foundation for balancing free expression with the need to address harmful misinformation.

Previous Attempts to Solve the Issue

EU Code of Practice on Disinformation (2018–Present)

This voluntary agreement between the European Commission and major tech platforms aims to tackle online disinformation through transparency, reporting, and fact-checking. However, enforcement has been limited and uneven.

Partnership on AI's Deepfake Detection Projects (2019–Present)

An initiative involving major tech companies like Microsoft and Meta, aiming to develop detection tools and public awareness. It has seen some technical success but lacks global reach and binding standards.

UNESCO Deepfake Literacy Campaign (2023)

Focused on building public awareness and resilience to synthetic media, this campaign highlighted the importance of education, but its impact is constrained by funding and scope.

Possible Solutions

Global Verification Protocols for Digital Media

Create a standardized, multilateral framework under the UN or ITU for authenticating and labeling digital content. Implement digital watermarks or cryptographic signatures for verified video/audio. Enforceable through international agreement and integration into platform algorithms.

AI Detection Infrastructure and Sharing

Fund and coordinate an open-source global AI detection infrastructure. Encourage data-sharing among states and tech companies to improve early warning systems for emerging deepfakes, especially during elections or crises.

Media Literacy Education and Awareness

Launch UN-led campaigns and regional partnerships to incorporate digital literacy and misinformation awareness into national education systems. Partner with local NGOs and educators to reach vulnerable populations.

Regulation of Generative AI Technologies

Promote international regulations on the development and deployment of generative AI tools. Require ethical impact assessments and clear usage disclosures by developers. Could be overseen by a new UN-affiliated tech ethics body.

Useful Links

United Nations - Artificial Intelligence and Global Governance: https://www.un.org/en

UNESCO - Recommendation on the Ethics of Artificial Intelligence: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Council on Foreign Relations - Deepfakes and Disinformation: https://www.cfr.org/

UN Resolution A/RES/76/227 - Countering Disinformation: https://docs.un.org/en/A/RES/76/227

OHCHR - Call for Inputs on Countering Disinformation:
https://www.ohchr.org/en/calls-for-input/2022/call-inputs-countering-disinformation-promotion-and-protection-human-rights

UN General Assembly Digital Library - Resolution A/RES/76/227:
https://digitallibrary.un.org/record/3955093?ln=en

UNESCO – "Internet for Trust" Initiative:
https://www.gov.br/g20/en/news/unesco-offers-recommendations-for-regulation-and-national-policies-on-ai

Bibliography

CONFERENCE REPORT DEEPFAKES, TRUST & INTERNATIONAL SECURITY.

"Document Viewer." *Un.org*, 2025, docs.un.org/en/A/RES/76/227.

https://apnews.com/author/david-klepper. "Scammers Are Impersonating CEOs and Trump Officials Using AI Deepfakes." *AP News*, 28 July 2025, apnews.com/article/artificial-intelligence-deepfake-trump-espionage-hack-scammers-da90ad1e5298a9ce50c997458d6aa610.

Miotti, Andrea, and Akash Wasil. "Combatting Deepfakes: Policies to Address National Security Threats and Rights Violations." *ArXiv.org*, 19 Feb. 2024, arxiv.org/abs/2402.09581.

"OHCHR | Call for Inputs: Countering Disinformation for the Promotion and Protection of Human Rights and Fundamental Freedoms." *OHCHR*, www.ohchr.org/en/calls-for-input/2022/call-inputs-countering-disinformation-promotion-and-protection-human-rights.

Poidevin, Olivia Le. "UN Report Urges Stronger Measures to Detect AI-Driven Deepfakes." *Reuters*, 11 July 2025, www.reuters.com/business/un-report-urges-stronger-measures-detect-ai-driven-deepfakes-2025-07-11/.

"The 2021 Innovations Dialogue: Deepfakes, Trust and International Security → UNIDIR." *Unidir.org*, 25 Aug. 2021, unidir.org/event/the-2021-innovations-dialogue-deepfakes-trust-and-international-security/.

UNESCO. "Ethics of Artificial Intelligence." *UNESCO*, 2023, www.unesco.org/en/artificial-intelligence/recommendation-ethics.

.