IPRI Islamabad Policy Research Institute
Research | Innovation | Dialogue | Policy

Policy Brief

Digital Deception in Warfare: An International Law Perspective on the Use of Deepfakes

Iqra Bano Sohail

March 2025

Chair International Law

Executive Summary

The rapid advancement of deepfake technology has introduced unprecedented challenges in modern warfare, particularly in the realms of deception, disinformation, and psychological operations. Utilizing Artificial Intelligence (AI) and machine learning, deepfakes generate hyper-realistic yet fabricated digital content, making them a powerful tool in military and geopolitical conflicts. Their misuse raises critical concerns under International Humanitarian Law (IHL), particularly regarding civilian protection, the principle of distinction, and the prohibition of perfidy. Recent conflicts have demonstrated the alarming weaponization of deepfakes, including their use to manipulate battlefield decisions, spread false surrender orders, and mislead civilian populations into life-threatening situations. The blurred line between lawful ruses of war and perfidious deception necessitates urgent legal clarity. While existing IHL provisions address deception in armed conflict, they do not explicitly regulate AI-driven misinformation tactics, creating legal loopholes that can be exploited.

Policy Recommendations

- The UN and international legal bodies should define deepfakes to ensure legal clarity, regulatory consistency, and accountability. This definition must cover all AI-generated or manipulated digital media, images, audio, and video, including fully synthetic and altered content. It should also highlight deepfake's deceptive nature and their potential to mislead audiences.
- The International Committee of the Red Cross (ICRC) should issue official commentaries explicitly prohibiting the use of deepfakes in armed conflict. These commentaries would clarify their legal status under IHL, reinforce humanitarian principles, and guide states in adapting national frameworks to address digital deception in warfare.
- The ICRC should provide detailed guidance on differentiating between lawful military deception and perfidious use of deepfakes in warfare. A dedicated commentary should outline scenarios where deepfakes constitute perfidy, such as fabricated surrender orders, and those that qualify as lawful ruses, such as misleading enemy troop movements.
- International Telecommunication Union (ITU) and the World Intellectual Property Organization (WIPO) should work towards standardizing AI-driven authentication mechanisms to verify media authenticity. This includes embedding digital markers, such as watermarks, at the moment of capture to enable continuous verification.

- To operationalize reversing the presumption of correctness, Standard Operating Procedures (SOPs) must be established for military personnel. They should be trained to disregard all digital communications unless they are explicitly confirmed through secure channels within their chain of command.
- Social media companies should integrate automated detection tools that flag and label potential deepfakes, utilizing AI-driven forensic analysis and metadata verification techniques.

Introduction

Deepfake technology represents a highly advanced form of synthetic media that utilizes Artificial Intelligence (AI) and machine learning to generate hyper-realistic yet deceptive digital content. While there is no universally accepted definition, deepfakes broadly encompass the creation or modification of images, videos, and audio recordings in a manner that convincingly replicates real individuals, making the fabricated content nearly indistinguishable from authentic media[1].

At the core of deepfake technology lies Generative Adversarial Networks (GANs), an AI framework comprising of two competing neural networks: a generator, which synthesizes artificial content, and a discriminator, which assesses its authenticity[2]. Through repeated iterations, the generator progressively improves the realism of its output until it becomes nearly indistinguishable from genuine media. This continuous refinement has significantly enhanced the quality of digital manipulation, making deepfakes increasingly difficult to detect, even with advanced forensic tools.

The rapid evolution of deepfake technology has given rise to serious legal, ethical, and societal challenges. It poses significant risks in areas such as governance, national security, individual privacy, and democratic integrity, necessitating urgent discussions on regulatory frameworks and mitigation strategies to address its potential misuse.

Military Applications of Deepfake Technology

Deepfake technology has emerged as a multifaceted tool in military operations, with various applications. The following categories encapsulate the primary uses of deepfakes in military contexts, highlighting their strategic implications.

Operational Deception in Warfare

This technology has been utilized as a tool for operational deception, as demonstrated by the fabricated video of Ukrainian President Volodymyr Zelenskyy falsely urging his troops to surrender[3]. This digitally manipulated content was designed to mislead enemy forces and

[1] Proofpoint 'What Is a Deepfake? Definition & Technology' https://www.proofpoint.com/us/threat-reference/deepfake accessed 1 February 2025.

[2] TechTarget 'What is Deepfake Technology?' https://www.techtarget.com/whatis/definition/deepfake accessed 1 February 2025.

[3] Center for Strategic & International Studies, Crossing the Deepfake Rubicon (11 January 2024) https://www.csis.org/analysis/crossing-deepfake-rubicon accessed 2 February 2025.

sow confusion among Ukrainian soldiers. By leveraging such disinformation, military actors can disrupt decision-making processes, potentially gaining a tactical advantage. This example highlights the growing role of psychological warfare in modern conflicts, where misinformation can swiftly influence battlefield dynamics and undermine the morale and cohesion of armed forces.

Endangering Civilians through Manipulation and Misinformation

In modern conflicts, deepfake technology has been increasingly weaponised to manipulate civilian populations, often with grave consequences. The dissemination of false content can be strategically employed to portray a misleading narrative, such as a false announcement of a ceasefire or assurances of safety[4]. This tactic may be employed with the intent to induce civilians to leave safe zones and return to areas that are still under active threat. By deliberately misleading non-combatants into life-threatening situations, such deceptive tactics not only endanger civilian lives but also raise profound ethical and legal questions regarding accountability and the protection of civilians in armed conflict.

Disinformation and Propaganda

Deepfakes have emerged as a powerful tool for disinformation and propaganda in military and geopolitical contexts, highlighting their role in modern information warfare. The ability to manipulate public perception through fabricated digital content has become a key strategy for state and non-state actors alike. Russia, in particular, has integrated deepfakes and advanced disinformation tactics into its broader information warfare strategy, aiming to manipulate public opinion and destabilize political systems, particularly in Europe.[5]

One of the most notable tactics employed by Russia involves the creation of cloned websites designed to spread misleading narratives. The "DoppelGänger"[6] campaign serves as a prominent example, where Russian actors developed counterfeit versions of reputable media outlets such as Bild and The Guardian to disseminate pro-Russian propaganda. By mimicking the appearance and branding of legitimate news platforms, these fake websites infiltrate the media landscape, lending credibility to false narratives and misleading the public. Such

[4] Lieber Institute for Law and Warfare “DEEPFAKES" AND THE LAW OF ARMED CONFLICT: ARE THEY A VIOLATION? https://lieber.westpoint.edu/deepfakes/ accessed 2 February 2025

[5] Foreign Policy *Information Warfare in Russia's War in Ukraine* https://foreignpolicy.com/2022/08/22/information-warfare-in-russias-war-in-ukraine/ accessed 4 February 2025

[6] Adam Majchrzak 'Russian disinformation and the use of images generated by artificial intelligence (deepfake) in the first year of the invasion of Ukraine' [Media Biznes Kultura nr 1 (14) Vol. 1 (14) 2023].

incidents underscore the growing threat posed by deepfake-enabled disinformation, necessitating stronger regulatory and technological measures to counter its impact.

International Law on The Use of Deepfakes in Armed Conflict

Although International Humanitarian Law (IHL) does not explicitly address the use of deepfakes, their application in specific contexts, such as acts of perfidy and psychological operations targeting civilians, raises significant legal and ethical concerns. The deliberate use of deceptive digital content in armed conflict can undermine fundamental IHL principles, including the protection of civilians, the prohibition of treachery, and the requirement of distinction between combatants and non-combatants.

The Legality of Deepfakes in Warfare: Distinguishing between Ruses and Perfidy

Under IHL, deception in warfare is not inherently unlawful, as military forces often employ strategic misinformation to gain an operational advantage[7]. However, IHL distinguishes between lawful ruses and prohibited acts of perfidy. Article 37(1) of Additional Protocol I (AP I) to the Geneva Conventions defines perfidy as

> *"acts inviting the confidence of an adversary to lead him to believe that he is entitled to, or is obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence."*

Perfidy is expressly forbidden when it results in death, injury, or capture. Classic examples include feigning surrender to lure an enemy into a vulnerable position before attacking, pretending to be a civilian while preparing an assault, or misusing protected emblems such as the Red Cross or Red Crescent to shield military operations[8].

In this context, deepfakes that fabricate an enemy commander's order to surrender, causing opposing forces to lower their defenses before being attacked, would constitute a clear violation of this provision. Such manipulative use of digital deception exploits the adversary's reliance on the laws of armed conflict and is, therefore, unlawful.

Conversely, not all uses of deepfake technology in warfare are prohibited. For example, generating deepfake content to mislead enemy troops about strategic deployments or military

[7] International Committee of the Red Cross *Ruse of War* https://casebook.icrc.org/a_to_z/glossary/ruse-war accessed 9 February 2025.

[8] ICRC, 'Ruse of War' (ICRC Resource Centre) https://casebook.icrc.org/a_to_z/glossary/ruse-war accessed 3 February 2025.

movements could fall within the category of lawful ruses of war. Similarly, falsified audiovisual materials presenting incorrect intelligence information may influence military decision-making without necessarily breaching international law. The legality of such actions depends on their adherence to IHL principles, particularly those ensuring the protection of civilians and prohibiting perfidious conduct.

Deepfakes and Civilian Protection under IHL

While certain military applications of deepfake technology may be legally permissible, their use against civilian populations raises serious concerns under IHL. Two key provisions of AP I to the Geneva Conventions establish clear limitations on such practices. Firstly, Article 57(1) encapsulates

> *"In the conduct of military operations, constant care shall be taken to spare the civilian population, civilians and civilian objects."*

Additionally, Article 51(2) entails that

> *"The civilian population as such, as well as individual civilians, shall not be the object of attack. Acts or threats of violence the primary purpose of which is to spread terror among the civilian population are prohibited."*

The deliberate use of deepfakes to incite fear and panic, particularly when disseminated through public platforms such as social media, can have severe consequences. Fabricated content falsely depicting an imminent nuclear strike, a large-scale natural disaster, or a fake surrender order could trigger mass hysteria, resulting in widespread chaos, civilian displacement, and even loss of life. In such cases, the deployment of deepfakes may violate the obligation to protect civilians from the effects of armed conflict under Article 57(1).[9]

Furthermore, if deepfake-generated disinformation is used with the intent to spread terror, such as falsely announcing an impending military attack or portraying fabricated acts of violence, it could constitute a violation of Article 51(2). Additionally, if such disinformation leads civilians to inadvertently expose themselves to hostilities, such as through false evacuation orders directing them into active conflict zones the responsible party could be held

[9] B Boutin, 'State Responsibility in Relation to Military Applications of Artificial Intelligence' (2022) 35 *Leiden Journal of International Law* 689.

liable for failing to uphold the principle of constant care[10]. The misuse of deepfakes in this manner not only undermines the protection of non-combatants but also challenges the fundamental humanitarian principles that govern armed conflict.

Options to Regulate the Use of Deepfakes

The Need for a Standardized Definition

A universally accepted definition of deepfake is crucial for ensuring legal clarity, regulatory consistency, and international cooperation in addressing its misuse. A standardized definition would enable legal systems to develop effective frameworks, promote accountability, and facilitate coordinated responses to emerging threats posed by this technology.[11]

One notable definition is provided in the European Union (EU) AI Act, which defines deepfakes as

> *"AI-generated or manipulated image, audio, or video content that resembles existing persons, objects, places, or other entities or events and would falsely appear to a person to be authentic or truthful."*

To establish a globally recognized definition, several key elements must be considered. First, the definition should encompass all forms of digital media, including images, audio, and video, that are generated or altered using artificial intelligence. Second, it must account for both entirely synthetic content and manipulated versions of existing media. Finally, it should emphasize the deceptive nature of deepfakes highlighting their ability to mislead individuals into perceiving fabricated content as authentic.

The establishment of a standardized definition will offer several benefits. It ensures regulatory consistency, allowing legal frameworks across jurisdictions to align, thereby preventing loopholes that could be exploited by malicious actors. It enhances international cooperation, enabling states and legal bodies to collaborate more effectively in addressing deepfake-related challenges. Moreover, it provides legal clarity, creating a well-defined basis for determining liability and enforcing accountability in cases of deepfake misuse.

[10] ICRC, International Humanitarian Law and the Challenges of Contemporary Armed Conflicts (ICRC, 2019)

[11] T Ramluckan, 'Deepfakes: The Legal Implications' (2024)

ICRC Commentaries on the Use of Deepfakes in Warfare

Prohibition on the Use of Deepfakes

As the guardian of the Geneva Conventions, the International Committee of the Red Cross (ICRC) plays a crucial role in ensuring that IHL remains relevant in the face of evolving threats. Given the increasing sophistication of deepfake technology and its potential impact on armed conflict, the ICRC should issue commentaries advocating for clear and explicit prohibitions on the use of deepfakes in warfare. These commentaries would provide essential legal clarity, reinforce humanitarian principles, and guide states in adapting their national frameworks to address this emerging challenge[12].

Currently, the Geneva Conventions do not explicitly address deepfake technology, creating gaps in the legal framework that could be exploited. Beyond deterrence, ICRC commentaries would contribute to the development of a strong international norm against the weaponisation of digital deception, reinforcing the fundamental principle that warfare must not involve manipulative tactics that distort reality and endanger non-combatants.

There is historical precedent for the ICRC using commentaries to address evolving forms of warfare. One such approach includes the ICRC's Interpretive Guidance on the Notion of Direct Participation in Hostilities (2009). This commentary addressed the complex issue of civilian participation in hostilities, a gray area in IHL, by clarifying when a civilian transitions from a protected person to a legitimate military target. This situation is analogous to the issue of deepfakes as existing laws on perfidy and deception require interpretation in the context of new technology.

Difference between Perfidy and Ruse

In addition, the ICRC can play a crucial role in providing further guidance on the distinction between lawful ruses and perfidious acts involving deepfakes[13]. A detailed commentary could offer concrete examples of deepfakes that would constitute perfidy, such as fabricated surrender orders designed to mislead enemy forces into lowering their defenses, and those that would be considered lawful ruses, such as deepfakes used to manipulate enemy troop movements without violating legal protections. The commentary could also clarify the essential

[12] Ibid 10

[13] Eric Talbot Jensen and Summer Crocket, 'Deepfakes' (West Point Lieber Institute for Ethics in Public Life, 2020).

elements of perfidy in the context of deepfakes, including the requirement of "inviting the confidence of an adversary" and the "intent to betray that confidence."

AI-Driven Authentication Mechanisms

As deepfake technology becomes increasingly sophisticated, proactive measures to verify the authenticity of images and videos are essential. One promising solution involves embedding digital markers, like watermarks, into media content at the moment of capture.[14] These AI-generated markers are highly sensitive to alterations, ensuring that any manipulation disrupts them, thereby facilitating the detection of tampering. Unlike traditional verification methods that assess authenticity only after content has been disseminated, this approach enables continuous verification throughout the media's lifecycle.[15]

Studies indicate that this method significantly enhances the accuracy of detecting manipulations while preserving the original content's quality. Furthermore, it is designed to withstand modifications commonly applied by social media platforms, ensuring the integrity of authenticity markers remains intact. To bolster global efforts against deepfakes, international organizations such as the International Telecommunication Union (ITU) and the World Intellectual Property Organization (WIPO) should explore the standardization of these digital markers.

The Role of Social Media Platforms in Deepfake Detection

Given the widespread dissemination of deepfakes via social media, platforms must play a proactive role in identifying and highlighting manipulated content. Social media companies should integrate automated detection tools that flag and label potential deepfakes, utilizing AI-driven forensic analysis and metadata verification techniques.[16]

Furthermore, social media platforms could collaborate with international regulatory bodies to establish transparency standards, ensuring that users are informed when consuming AI-generated content. The incorporation of "deepfake disclaimers" or "verified authenticity

[14] H Bao, X Zhang, Q Wang, K Liang, Z Wang, S Ji, and W Chen, 'Fake It Till You Make It: Realistic Deepfake Video Detection via Spatiotemporal Consistency Learning' (Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, 2024) 4122.

[15] Y Zhao, B Liu, M Ding, B Liu, T Zhu and X Yu, 'Proactive Deepfake Defence via Identity Watermarking' (2023) IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 4591.

[16] A Naitali et al, 'Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions' (2023) 12(10) Computers 216 https://doi.org/10.3390/computers12100216

badges"[17] on posts containing manipulated media could serve as an early warning mechanism, helping mitigate the spread of deceptive content.

Operationalizing the Presumption of Falsity Doctrine (Reversing the Presumption of Correctness)

Digital communications are often assumed to be truthful by default.[18] This presumption of correctness can facilitate the spread of misinformation, particularly in high-stakes environments such as armed conflict. To mitigate this risk, it is necessary to reverse this assumption. Instead of presuming digital messages to be accurate, they should be treated as false unless verified through reliable means.

To operationalize this shift, Standard Operating Procedures (SOPs) must be established for military personnel[19]. They should be trained to disregard all digital communications unless they are explicitly confirmed through secure channels within their chain of command.

Institutionalizing this norm within IHL requires the involvement of the International ICRC. Given its role in shaping IHL principles, the ICRC is well positioned to initiate discussions and establish guidelines for reversing the presumption of correctness in digital communications.

A formal resolution by UNGA could further codify this approach, providing clear guidelines for its implementation in armed conflict settings. By addressing the risks posed by misinformation, such a resolution would help ensure consistent application across nations and reinforce global efforts to maintain information integrity in warfare.

Policy Recommendations

- The UN and international legal bodies should define deepfakes to ensure legal clarity, regulatory consistency, and accountability. This definition must cover all AI-generated or manipulated digital media, images, audio, and video, including fully synthetic and altered content. It should also highlight deepfake's deceptive nature and their potential to mislead audiences.

[17] European Data Protection Supervisor, 'Deepfake detection' (*EDPS*, 2025)

[18] L Safrai and M Wellerstein, 'Liar's War: Protecting Civilians from Disinformation During Armed Conflict' (2023) 95(914) International Review of the Red Cross 475

[19] S Kehrt, 'All Warfare Is Based on Deception—Troops, Vets Targeted by Disinformation Can Fight Back' (The War Horse, 8 September 2022)

- ICRC should issue official commentaries explicitly prohibiting the use of deepfakes in armed conflict. These commentaries would clarify their legal status under IHL, reinforce humanitarian principles, and guide states in adapting national frameworks to address digital deception in warfare.
- The ICRC should provide detailed guidance on differentiating between lawful military deception and perfidious use of deepfakes in warfare. A dedicated commentary should outline scenarios where deepfakes constitute perfidy, such as fabricated surrender orders, and those that qualify as lawful ruses, such as misleading enemy troop movements.
- ITU and WIPO should work towards standardizing AI-driven authentication mechanisms to verify media authenticity. This includes embedding digital markers, such as watermarks, at the moment of capture to enable continuous verification.
- To operationalize reversing the presumption of correctness, Standard Operating Procedures (SOPs) must be established for military personnel. They should be trained to disregard all digital communications unless they are explicitly confirmed through secure channels within their chain of command.
- Social media companies should integrate automated detection tools that flag and label potential deepfakes, utilizing AI-driven forensic analysis and metadata verification techniques.

Action Matrix

Options for International Community

Option	Pathways to Solution	Implementation of Solution	Actors Responsible	Implementation Timelines
The Need for a Standardized Definition	A universally accepted definition of deepfake technology is essential for ensuring legal clarity, regulatory consistency, and international cooperation. Without a standardized definition, addressing the threats posed by deepfakes remains inconsistent and ineffective across jurisdictions.	Convene global discussions under UN bodies, such as the UN General Assembly (UNGA) and the UN Interregional Crime and Justice Research Institute (UNICRI), to develop a comprehensive and legally recognized definition.	• United Nations General Assembly • United Nations Interregional Crime and Justice Research Institute	3-6 Months for Initial consultations among international organizations and legal experts to draft a standardized definition. 3-6 Months for multilateral negotiations to refine and finalize the definition. 6-12 Months for adoption of the definition through international agreements or resolutions.
ICRC Commentaries on the Use of Deepfakes in Warfare	The ICRC can issue official commentaries clarifying the prohibition of deepfakes under IHL, reinforcing legal principles, and providing guidance on distinguishing between lawful ruses and perfidious acts.	The ICRC will conduct legal assessments, engage with international legal experts, and publish authoritative commentaries to guide states and military institutions.	• International Committee of the Red Cross • UN Office of the High Commissioner for Human Rights	12-18 Months for drafting, review, publication and dissemination of commentaries
AI-Driven Authentication Mechanisms	Develop and implement standardized digital watermarking and AI-driven verification systems to authenticate media content at the moment of capture.	International organizations and tech firms collaborate to create global standards, integrate authentication mechanisms into digital platforms, and ensure regulatory compliance.	• International Telecommunication Union • World Intellectual Property Organization • Social Media Platforms	18-24 Months for standardization discussions, regulatory frameworks, industry adoption and integration into media platforms