NSM#Blog, ed. Ursula Werther-Pietsch and Friedarike Santner

Quote: Dolezal, AI and Multilateralism,
NSM#Blog October 2025, https://nsm.uni-graz.at/en/the-nsmblog

NEW SCHOOL OF MULTILATERALISM
NSM#BLOG

September 2025

AI and Multilateralism – A Social-Ethical Perspective of the Dangers of Deepfakes and Disinformation

By Eugen Dolezal

Introduction

The digital transformation enables new forms of interaction, but also raises ethical questions. In addition to many opportunities, artificial intelligence in particular, or, in Kirchschläger's nomenclature, *data-based systems* (University of Lucerne, 2024), is contributing to rapid change in communication and power relations. Highly realistic "deepfake" videos and systematic disinformation threaten democratic debate and social cohesion. At the same time, multilateralism appears more important than ever as a new dimension of global coexistence: only when states, international organizations, and civil society actors act together can global challenges be regulated and mitigated. This article aims to briefly examine the dangers of deepfakes and disinformation from a socio-ethical perspective and, based on Kirchschläger's work, tries to build a bridge to multilateralism.

Definition of terms: Data-based systems (DS) and disinformation

"Artificial intelligence" as data-based systems

Many debates about "AI" confuse algorithmic systems with the idea of human-like intelligence, starting with the very term used to describe these systems. Kirchschläger criticizes this confusion of terms and recommends referring to data-based systems (DS), because these "AIs" are programmed by humans and are not autonomous. In general, it is more important to consider whether DS are based on human rights (Shamira & Kirchschläger, 2024) and how this can be guaranteed. This brings the ethical responsibility for the design, development, use, and dismantling of DS into focus.

Disinformation and deepfakes

The difference between misinformation and disinformation allows for a more nuanced assessment of the phenomenon, which is why it should be briefly introduced here. Misinformation refers to false information without malicious intent, while disinformation refers to deliberately false information intended to cause harm (Yamaoka-Enkerlin, 2020).

Between the two lies malinformation: true information used in a misleading context (Yamaoka-Enkerlin, 2020). Deepfakes are synthetic media (e.g., videos, audio, or images) generated using machine learning methods that appear "hyperrealistic" and undermine trust in digital content (Anand & Bianco, 2021). They exploit the fact that people perceive audiovisual impressions as authentic and can generate deceptively real actions or statements by individuals (Anand & Bianco, 2021).

Deepfakes and disinformation as a socio-ethical challenge

A UNIDIR study on deepfakes reminds us that trust is central to international security and social cohesion (Anand & Bianco, 2021). Deepfakes make it difficult to distinguish between truth and falsehood; the so-called *Liar's Dividend* means that even genuine recordings can be suspected of being manipulated, as the authenticity of audiovisual content can be questioned due to the mere possibility of being confronted with a deepfake (Yamaoka-Enkerlin, 2020, p. 731). We must warn of the far-reaching effects of this phenomenon, as easily accessible tools can amplify the spread of disinformation and the erosion of truth[1] (Anand & Bianco, 2021). The danger is not only political in nature. In the areas of deepfake pornography and digital identity theft, deepfake applications now enable even people with little technical expertise to create synthetic digital content (Increasing Threats of Deepfake Identities, n.d.). It should be noted that disinformation spread through deepfakes can, in principle, damage reputations and affect the dignity of those affected, especially since such deepfakes are also used to expose and delegitimize political opponents (Yamaoka-Enkerlin, 2020, pp. 731–732). This already points to several socio-ethical problems, which will only be mentioned here in a simplified and incomplete manner:

Human/personal dignity: It is striking that those affected are degraded to objects of manipulative strategies and that their expression in the world is taken away from them. In addition, deepfakes can be used to generate content that results in social ostracism and exclusion for those affected.

Common good and democracy: False or misleading information undermines public deliberation and opinion-forming processes and can thus influence elections and decision-making. It is particularly problematic that anonymous dissemination not only sows

[1] For more detailed discussion of the concept of truth within the digital sphere, see (Filipović, 2024).

fundamental mistrust in media content, but also delegitimizes authentic content in principle through the liar's dividend.[2]

Justice: Since the concept of justice is extremely broad, only a brief list of issues will be mentioned here, which could definitely be expanded upon, but at the same time overlaps with what has already been discussed. As a multidimensional challenge to justice, deepfakes violate epistemic justice by distorting truth and access to information; they undermine relational justice by eroding trust and social cohesion; and, as a technological application with communicative and political power, they are unevenly distributed in terms of their power potential despite their ease of use.[3] In addition, deficits in terms of restorative justice are to be expected because responsibility and redress are difficult to assign. Finally, the global nature of digital disinformation calls for structural justice, something that requires multilateral cooperation and human rights-based regulation.

In the light of this supranational relevance, the following section briefly outlines an approach developed by Peter Kirchschläger that addresses the issue and necessity of multilateral regulation and oversight in the context of DS.

Human rights-based regulation for DS

Since only humans have conscience, freedom, and moral judgment, and DS, as programmed entities, act solely heteronomously (Kirchschläger, 2022, pp. 486–487), ethical responsibility for such systems cannot be delegated. Nor can it be delegated to so-called

[2] It is also noteworthy that, at least in the context of academic publications, transparency regarding the use of generative DS does not contribute to strengthening trust; on the contrary, trust in the author declines when this use is made transparent (Schilke & Reimann, 2025). In this respect, there is an incentive not to disclose such use, as DS are apparently considered fundamentally untrustworthy in this context. Nevertheless, content that is not recognized as generated is perceived as convincing, a problem that is also evident in the context of disinformation through deepfakes.

[3] Hier soll klar differenziert werden, dass es an dieser Stelle nicht darum geht die technologischen Anwendung zur Erstellung von Deepfakes noch breiterer zugänglich zu machen, sondern darauf hingewiesen werden soll, dass trotz der prinzipiell niedrigen technischen Hürde der Erstellung dieser synthetischen Medieninhalte, die machtpolitische Nutzung von Deepfakes zur Desinformation, ihre zielgerichtete Verbreitung und das Präsenthalten selbiger im öffentlichen Diskurs mit bereits im Vorfeld vorhandener ökonomischer, politischer und struktureller Macht korreliert. Dies betrifft natürlich in besonderer Weise politische Akteure und Staaten aber auch Plattformen, welche durch ihre Algorithmik auf die Verbreitung von spezifischen politischen Inhalten im generellen und Deepfakes im speziellen Einfluss nehmen können.

"moral technologies," because even if data-based systems follow moral rules, they do so "without knowledge of the ethical quality of the rules" (ibid, p. 488).

On this basis, Kirchschläger calls for consistently human rights-based data systems (HRBDS). Human rights function as universal ethical reference points based on the shared vulnerability of human beings: those who recognize their own vulnerability must grant themselves and others equal rights to protection (Kirchschläger, 2022, p. 491). Kirchschläger makes a clear distinction between legitimate and illegitimate technological action: systems are ethically illegitimate if they violate human rights, for example through surveillance, discrimination, or disinformation (Kirchschläger, 2022, pp. 488–489). Conversely, "technology-based progress and ethics are not contradictory if data-based systems are designed in accordance with human rights" (ibid, 2022, p. 493).

In this respect, a fundamental normative reversal is required from an ethical point of view: it is not people who should adapt to technology, but technology that should adapt to human rights.

Multilateral approaches

Starting from the diagnosis that DS pose existential risks, not only because of their role in generating and spreading disinformation, but also because of algorithmic discrimination or as autonomous weapon systems, the question arises as to what extent the ethical principle described above can be transferred to the level of international politics, especially since DS can only be controlled to a limited extent at the national level due to their global operations. Against the global backdrop of the challenge posed by DS, its regulation requires an institutional form of global responsibility.

Kirchschläger proposes the establishment of an "International Data-Based Systems Agency (IDA)" to operate in a similar way to the United Nations' International Atomic Energy Agency (IAEA) (Shamira & Kirchschläger, 2024, pp. 4–6). This IDA would be based on three pillars:

1. Ethics-based regulation: international standards based on human rights.
2. Transparency and oversight: monitoring and assessment of risks posed by data-based technologies.
3. Knowledge justice: fair distribution of technical knowledge and access to secure technology.

The aim of this institution is to make these ethical principles binding through multilateral agreements. The intention is to combine ethical reflection with governance structures that go beyond voluntary commitments and thus make them internationally enforceable.

Kirchschläger's reference to the IAEA model is no coincidence: just as the use of nuclear energy required an international architecture of trust, the new challenges posed by DS also require a trust-building "moral infrastructure" that institutionally safeguards transparency, responsibility, and accountability (Shamira & Kirchschläger, 2024, p. 9).

However, this implies a specific understanding of multilateralism that goes beyond mere political cooperation and is structurally committed to ethics: a just digital order can only emerge when states limit their technological sovereignty through joint moral self-commitment. Multilateralism thus becomes the locus of normative containment of technical power.

Efforts such as the IDA proposed by Kirchschläger do not arise in a vacuum. For example, the UN has been working for several years to create a coherent global framework for DS, on the one hand in view of the risks already mentioned, but on the other hand also because 164 of 193 UN states do not have their own "AI" strategy and governance therefore remains fragmented or non-existent (Fournier-Tombs & Siddiqui, 2024). In this respect, it is hardly surprising that, in addition to efforts such as those of the IDA, the UN Secretariat is also attempting to bring momentum to the issue of DS governance at the multilateral level, with the Global Digital Compact being presented in September 2024. This compact aims to formulate common principles for an open, free, and secure digital future and to help advance governance in the context of DS (Fournier-Tombs & Siddiqui, 2024).

Conclusion

The danger of deepfakes and disinformation shows how vulnerable democracy and human rights are in the digital age. From a social ethics perspective, this forces us to reflect on existing power relations, to remember that human dignity and the common good must be at the center, and that we are obliged to act in solidarity with one another. "AI" is not magical or even superhuman intelligence, but a data-based system programmed by humans that can be regulated, as the AI Act shows, and should also be regulated globally. Multilateralism is indispensable in this regard, as individual states alone can hardly or not at all exercise control over global platforms. The UN discussion on the Global Digital

Compact and other documents such as the UNESCO Recommendation on the Ethics of Artificial Intelligence show that a new consensus is possible. The idea of creating an international agency for data-based systems could be worth pursuing in this regard.

***Eugen Dolezal**, Pre Doctoral Fellow, Magister of Social Ethics and Systemic Theology, University of Graz. "AI and Human Rights", Vol. 6 NSM Book Series by Nomos, tbp 2027.*

Literature

Anand, A., & Bianco, B. (2021). *Deepfakes, Trust & International Security. Conference Report* (United Nations Institute for Disarmament Research (UNIDIR), Ed.). UNIDIR. https://unidir.org/files/2021-12/UNIDIR_2021_Innovations_Dialogue.pdf

Filipović, A. (2024). Wahrheit. In *Digitale Ethik*. Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783748942399

Fournier-Tombs, E., & Siddiqui, M. (2024). Wie kann künstliche Intelligenz global gesteuert werden? *Vereinte Nationen*, *72*(5), 195–201. https://doi.org/10.35998/vn-2024-0021

Increasing Threats of Deepfake Identities. (n.d.). U.S. Department of Homeland Security.

Kirchschläger, P. G. (2022). Ethische KI? Datenbasierte Systeme (DS) mit Ethik. *HMD Praxis Der Wirtschaftsinformatik*, *59*(2), 482–494. https://doi.org/10.1365/s40702-022-00843-2

Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, *188*, 104405. https://doi.org/10.1016/j.obhdp.2025.104405

Shamira, A., & Kirchschläger, P. G. (2024). *Governing Global Existential AI Risks: Lessons from the Intl. Atomic Energy Agency* (Inclusive Digital Transformation, p. 18) [Policy Brief]. G20 Brasil 2024. https://www.t20brasil.org/media/documentos/arquivos/TF05_ST_05_GOVERNING_GLOBAL_EX66d7093af049f.pdf

University of Lucerne. (2024, September 24). *Artificial intelligence: Peter G. Kirchschläger advises G20*. University of Lucerne. https://www.unilu.ch/en/news/artificial-intelligence-peter-g-kirchschlaeger-advises-g20-8808/

Yamaoka-Enkerlin, A. (2020). Disrupting Disinformation: Deepfakes and the Law. *New York University Journal of Legislation & Public Policy*, *22*(3), 725–749.