Rachel Ntow

December 4th, 2025

The Deepfake Blindspot in AI Governance

Deepfakes are no longer just a novelty or disinformation threat — they're reshaping digital trust. Despite mounting evidence of their harms in multiple domains, regulatory responses remain fragmented and misaligned with the actual threats. MSc Development Management candidate Rachel Ntow argues that this governance gap is systematically eroding the accountability and digital credibility that institutions rely on.

Recently, while scrolling through YouTube Shorts, I came across a viral TikTok video showing a US-certified doctor whose content I regularly follow, apparently giving medical advice. The problem? The doctor had never made the video. The health advice was fabricated, potentially harmful, and the synthetic clip was convincing enough to mislead thousands. When the real doctor tried to flag the video and contact the creator, he was blocked. With the video still circulating, most viewers may act on false information with real medical consequences.

This incident reflects an underlying issue in AI governance. Although researchers have extensively documented the growing power of deepfakes and the harms they cause, regulatory responses remain fragmented, reactive, and misaligned with how the technology is abused in practice. Deepfakes are not just a content-moderation issue or an election-season threat. They are a systemic risk multiplier: a technology that exploits digital authenticity to facilitate illicit activities such as financial fraud, undermine public health, violate personal autonomy, and erode public trust.

The Tangible Price of Deception and the Governance Gap

Deepfakes remain notably peripheral in governance frameworks, exposing how current AI regulation misdiagnoses the challenge. Legislative efforts, including the EU's Digital Services Act and AI Act largely treat deepfakes as a content distribution and disclosure problem solvable through platform accountability i.e. transparency, labelling, and platform moderation. Although relevant, this approach overlooks the speed at which deepfakes have become sophisticated tools of organized crime.

Regulatory debates fixate on election interference through misinformation and cyber violence such as non-consensual intimate imagery. Although important, this narrow lens leaves blind spots around economic and public safety risks. The widely reported case of the $25 million Arup executive impersonation scam in Hong Kong showed how voice cloning and synthetic video can defeat even robust safeguards. Deepfake-based identity fraud is now able to bypass common verification systems, including KYC checks. According to Deloitte, deepfake-enabled fraud has already resulted in significant losses, with estimates for the US market alone reaching $12.3 billion in 2023. These figures are expected to increase exponentially, reaching a projected $40 billion in the US by 2027, impacting consumers, businesses, and governments alike.

Legal frameworks compound the gap through conceptual inadequacy. Denmark's attempt to protect digital likeness under copyright law is commendable as a proactive step to safeguard people's rights, yet it also highlights the reactive nature of current approaches: copyright governs creative works, not a person's identity. This limitation underscores the absence of modern legal frameworks that recognize and protect digital identity in an age where anyone's face or voice can be cloned and misused.

The Credibility Crisis: Eroding Epistemic Trust

Beyond direct harm, deepfakes fuel an emerging crisis of credibility. Researchers call this the 'liar's dividend' — the ability to dismiss genuine evidence simply by claiming it's fake. The result is a serious erosion of accountability and a weakening of institutions that depend on verifiable proof.

This erosion of trust also has economic costs. Businesses, financial institutions, and public agencies now invest heavily in verification technologies simply to maintain basic operational confidence. Every digital interaction becomes slower, more expensive, and more complex. Deepfakes don't just deceive individuals; they degrade the trust infrastructure modern societies rely on.

Why the Blindspot Persists

Several structural issues hinder the translation of research insight into regulatory action. For instance, governance remains fragmented, with media law, cybersecurity policy, financial regulation, and AI governance operating in silos, thereby allowing malicious actors to exploit institutional gaps. Existing regulation is minimal and reactive; even where deepfakes are acknowledged, such as in the EU AI Act limited disclosure requirements, measures are narrow, and weakly enforced, offering little deterrence to perpetrators.

Meanwhile, legal adaptation lags behind ethical debate: although scholars have long examined questions of digital authenticity, consent, and epistemic harm, these discussions rarely materialize into enforceable protections. As a result, victims face unclear paths to remedy, and penalties for misuse remain inconsistent across jurisdictions.

Bridging the Gap: What Governance Must Address

A stronger governance framework needs to move beyond piecemeal solutions and reflect the real-world risks documented by researchers and industry experts. Several priorities stand out:

i. **A Risk-Based Classification for Deepfake Abuse** – Deepfakes used for financial fraud, impersonation, or large-scale misinformation should be classified as high-risk AI systems not limited risk as seen in the EU AI Act. Giving it the attention it deserves would trigger mandatory controls, audits, and accountability requirements, not voluntary guidelines.

ii. **Modern Digital Identity Rights** – Countries need clear legal protections for digital likeness i.e. voice, image, and biometric identity as independent rights. This ensures individuals have legal recourse when their identity is synthesized or exploited.

iii. **Cross-Sector and Cross-Border Coordination** – Because deepfake abuse cuts across industries and borders, governance must do the same. Technology companies, financial institutions, law enforcement, and regulators need shared detection tools, threat intelligence, and rapid-response channels. Voluntary partnerships are not enough.

iv. **Accessible, Victim-Centred Remedies** – Legal pathways must be simple, enforceable, and meaningful. Statutory damages should be strong enough to deter misuse, and remedies must account for the speed with which synthetic media spreads globally.

The Cost of Inaction

The deepfake challenge reveals a broader issue in AI governance: the significant lag between what researchers understand and what policymakers act upon. We already know what deepfakes can do, how they're misused, and what harms they cause. What is missing is coordinated political will.

If regulatory frameworks continue to treat deepfakes as isolated nuisances rather than structural threats, they will progressively weaken the digital trust systems that underpin economies, public safety, and accountability. The blind spot is not about a lack of knowledge; it is about failing to act on what we know. That failure allows synthetic deception to corrode institutional trust and democratic accountability.

The views expressed in this post are those of the authors and in no way reflect those of the International Development LSE blog or the London School of Economics and Political Science.

Featured image credit: Created by author in Adobe Express.

About the author

Rachel Ntow

Rachel Ntow is an MSc Development Management candidate at the London School of Economics, with a strong interest in digital inclusion and AI governance. With over four years of experience in international development, she has collaborated with development partners, policymakers, and key stakeholders to deliver seven international projects advancing inclusive education and human capital development. She is passionate about leveraging digital technology to drive social impact and inclusive growth.

Posted In: Digitalisation and ICT

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE