Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

SPOTLIGHT ON DEEPFAKES: MAPPING RESEARCH AND REGULATORY RESPONSES

Alena Birrer
University of Zurich

Natascha Just
University of Zurich

Background

In February 2024, an open letter with over a thousand signatures from tech executives, academics, and prominent figures from the entertainment industry was released, urging governments to take immediate action against the proliferation of deepfakes (*"Disrupting the Deepfake Supply Chain"*, 2024). This urgency for regulatory action stems from growing concerns about the threats posed by deepfakes to both individuals and society. Journalists have vividly painted a dystopian scenario, referring to deepfakes as an "epistemic apocalypse" (Habgood-Coote, 2023), where the lines between authentic and artificial content become alarmingly blurred, creating a sense of impending doom (Wahl-Jorgensen & Carlson, 2021). This is accompanied by fear-mongering by the AI industry through public warnings about the dangers of deepfakes (Bartz, 2023). At the same time, regulators have been noticeably more hesitant in their responses. The European Artificial Intelligence Act (AI Act), for instance, classifies deepfakes as "limited risk AI systems" and sets only minimal transparency requirements. The United States and China have implemented more targeted legislation, criminalizing the distribution of potentially harmful deepfakes. This was accompanied, however, by concerns that governments could use such rules to restrict free speech and control information flows (Hine & Floridi, 2022; Tsukayama et al., 2019).

Aim and Methodological Approach

Given this background, the questions arise: How much do we actually know and understand about deepfakes, and what regulatory responses have emerged in response? A growing research field discusses deepfakes' potential harms (e.g., Chesney & Citron, 2019; Diakopoulos & Johnson, 2021; Rini & Cohen, 2022) and

Suggested Citation (APA): Birrer, A., & Just, N. (2024, October). *Spotlight on Deepfakes: Mapping Research and Regulatory Responses.* Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org. Full paper available at https://doi.org/10.1177/146144482412531

whether existing public and private law is sufficient to counteract them (e.g., Caldera, 2019; Meskys et al., 2020; van der Sloot & Wagensveld, 2022). In contrast, there is a lack of consolidated knowledge regarding the empirical evidence supporting these concerns as well as the specific regulatory measures developed in response. To bridge these gaps, our methodological approach is two-fold: (1) we provide a systematic literature review to consolidate what is currently empirically known about deepfakes, and (2) a qualitative content analysis of the evolving regulatory landscape. This is to offer a more comprehensive understanding of the deepfake phenomenon and to provide directions for future research and policymaking.

Systematic Literature Review

We conducted a systematic literature review of empirical research on deepfakes to consolidate existing knowledge regarding their current uses, effects, consequences, and regulatory challenges. Relevant literature was identified following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021). This included a broad search of five databases (Scopus, Web of Science, EbscoHost, ProQuest, arXiv). A total of 79 studies was finally included in the analysis. Given the diversity of research questions and methods employed across the studies, we chose a qualitative approach based on a deductive-inductive coding scheme to review the literature. The findings suggest that our understanding of deepfakes remains insufficient to assess the validity of the often-voiced concerns about their negative effects and the most effective strategies to counteract them. The term “deepfake” is often not precisely defined and lacks clear conceptual boundaries, which further complicates efforts to grasp the phenomenon and its distinct manifestations. Additionally, empirical evidence on the prevalence of deepfakes is largely missing. The literature review shows that emphasis is on three major regulatory challenges (i.e., difficulties in detecting deepfakes, deepfake disinformation, and deepfake pornography). It further indicates that deepfakes currently do not introduce fundamentally new and unique regulatory challenges. Instead, they add to the repertoire of tools available for spreading harmful or illegal content such as disinformation and non-consensual pornography. Moreover, there is some evidence that common countermeasures such as raising awareness or labelling deepfake content may not be as effective in mitigating their potential harms. In some cases, these strategies can even backfire and result in a general climate of uncertainty and mistrust in media.

Qualitative Content Analysis of Regulatory Responses

Informed by the findings of the systematic literature review, we conducted an in-depth qualitative content analysis of dedicated regulatory responses to deepfakes. We identified 100 policy and legal documents through a review of existing research, in-depth searches of regulatory authorities’ websites, and thorough monitoring of media coverage and policy blogs. The documents were coded using the coding software MAXQDA for qualitative data analyses. Special attention was paid to examining and evaluating the rationales driving the need for regulatory action, identifying the accountable actors, and assessing the adequacy and efficacy of the proposed measures in the context of existing empirical research on deepfakes. Overall, a diverse spectrum of regulatory responses to deepfakes emerged, ranging from market-driven initiatives to state-imposed command-and-control-regulation, with various forms of self- and co-regulation in between. The measures address different stages of the deepfake

lifecycle and vary in their target, applying to producers of deepfake technology, users who create or disseminate deepfakes, or the platforms that host them. Some policymakers have chosen to refrain from regulatory action altogether, either due to limited research on deepfakes or the belief that existing laws – although sometimes extended in scope – are generally well-equipped to address deepfakes. Others rely on self- and co-regulation aimed at raising awareness as well as hard, i.e. legally binding and enforceable regulations that require transparency, or ban or preemptively curb the production or distribution of potentially harmful deepfakes. Such measures address the regulatory challenges described in the literature, focusing primarily on electoral manipulation through deepfake disinformation and deepfake pornography. In this context, they often refer to empirically unverified assumptions about the prevalence and deceptive capacity of deepfakes. Moreover, when evaluated against the findings of the literature review, concerns about enforcement and efficacy of the countermeasures persist.

Conclusion and Outlook

The dynamic nature of deepfake technology calls for adaptive policy approaches (Latzer, 2013) that aim to mitigate harm while protecting individual rights and addressing larger societal issues. Risk-based approaches, as proposed in the AI Act, appear to hold the most promise in striking this balance. Nonetheless, it is crucial to acknowledge that existing tools may not fully resolve current and future challenges, emphasizing the need for critical oversight and periodic review. This must also include careful consideration of adequate governance arrangements, including both appropriate state and private involvement as highlighted in the governance-choice approach (Latzer et al., 2019). Altogether, this highlights the necessity for further empirical research to navigate and comprehend the regulatory challenges raised by deepfakes. Future research should specifically clarify the conceptual boundaries and the diverse applications of deepfakes and explore the individual and societal impact they can have (both harmful and beneficial), especially beyond the Global North. In addition, more research is needed on the intended and unintended consequences of countermeasures to strengthen evidence-based policymaking.

References

Bartz, D. (2023, May 25). Microsoft chief says deep fakes are biggest AI concern. *Reuters*. https://www.reuters.com/technology/microsoft-chief-calls-humans-rule-ai-safeguard-critical-infrastructure-2023-05-25/

Caldera, E. (2019). “Reject the Evidence of Your Eyes and Ears”: Deepfakes and the Law of Virtual Replicants. *Seton Hall Law Review*, *50*(1), 177–205.

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media and Society*, *23*(7), 2072–2098. https://doi.org/10.1177/1461444820925811

Disrupting the Deepfake Supply Chain. (2024). https://openletter.net/l/disrupting-deepfakes

Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese*, *201*(3). https://doi.org/10.1007/s11229-023-04097-3

Hine, E., & Floridi, L. (2022). New deepfake regulations in China are a tool for social stability, but at what cost? *Nature machine intelligence*, *4*(7), 608–610. https://doi.org/10.1038/s42256-022-00513-4

Latzer, M. (2013) Medienwandel durch Innovation, Ko-Evolution und Komplexität. Ein Aufriss. *Medien & Kommunikationswissenschaft, 61*(2): 235–252. https://doi.org/10.5167/uzh-78012

Latzer, M., Saurwein, F., & Just, N. (2019). Assessing Policy II: Governance-Choice Method. In H. Van den Bulck, M. Puppis, K. Donders, & L. Van Audenhove (Eds.), *The Palgrave Handbook of Methods for Media Policy Research* (pp. 557–574). Springer International Publishing. https://doi.org/10.1007/978-3-030-16065-4_32

Meskys, E., Liaudanskas, A., Kalpokiene, J., & Jurcys, P. (2020). Regulating deep fakes: Legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, *15*(1), 24–31. https://doi.org/10.1093/jiplp/jpz167

Page, M. J., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*(71), 1–9. https://doi.org/10.1136/bmj.n71

Rini, R., & Cohen, L. (2022). Deepfakes, Deep Harms. *Journal of Ethics and Social Philosophy*, *22*(2), 143–161. https://doi.org/io.265s6/jesp.v22i2.162828

Tsukayama, H., McKinney, I., & Williams, J. (2019). *Congress Should Not Rush to Regulate Deepfakes*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2019/06/congress-should-not-rush-regulate-deepfakes

van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: Regulatory challenges for the synthetic society. *Computer Law and Security Review*, *46*. https://doi.org/10.1016/j.clsr.2022.105716

Wahl-Jorgensen, K., & Carlson, M. (2021). Conjecturing Fearful Futures: Journalistic Discourses on Deepfakes. *Journalism Practice*, *15*(6), 803–820. https://doi.org/10.1080/17512786.2021.1908838