The European AI Act and its shortcomings in safeguarding young individuals from deepfakes

Maximilian Weber

Identification of subject

Discussions on artificial intelligence are ubiquitous in the fields of business and information technology (IT). In this context, the European Union's adoption of the Artificial Intelligence Act in March 2024 marks a significant milestone, positioning the EU as the first to regulate AI for companies and governments. However, the EU AI Act exhibits notable gaps, particularly in protecting minors and adolescents against AI threats such as deepfake technology. To critically assess the law's effectiveness in protecting young people, I examine the law from technological, legal and business perspectives.

Personal motivation and rationale

Artificial intelligence has become highly influential due to its user-friendly nature and its ease for rapid generation of media content. My interest in AI therefore stems from its increasingly pervasive influence on our lives. The recent AI law passed by the EU Parliament is particularly significant for us Europeans, regulating AI applications such as facial recognition, CCTV and credit scoring (European Parliament, 2024). Despite the law's groundbreaking nature, surprisingly, both academia and EU legislation fail to address the protection of young people, a vulnerable group particularly susceptible to AI threats like deepfakes. Given its far-reaching consequences on young people's personal and professional futures, my research aims to address this oversight and advocate for enhanced protection measures for the EU's younger population.

Research question

My central question is to what extent the EU AI Act adequately protects young individuals from deepfake technology. The underlying hypothesis states that extending the AI law could oblige legal institutions and corporate businesses to introduce in-depth protection measures for young people and thereby implement effective, far-reaching protection mechanisms.

Literature review

Methodologically, my research builds on the latest literature in this field. However, the literature review indicates that the focus on protecting minors is a relatively new and unexplored aspect of the EU AI Act. The enacted EU AI law is based on the EU Commission's proposal from April 2021, which aimed to address the rapid advancement of AI systems and their challenges through consistent regulations across the EU (Ebers et al., 2021). Specifically, the legislators adopted a risk-based approach, categorising AI

systems into four risk levels: minimal, transparency, high and unacceptable risks (Madiega, 2024). The legal framework ratified in March 2024 now categorises an extensive range of AI systems into these levels, e.g. CCTV under high risk and deepfakes – curiously – only under transparency risk. In this regard, the AI law addresses previous criticisms of unclear definition of risk systems (Hacker, 2021) and thereby illustrates the EU Parliament's commitment to align the legislation with current research findings.

On the one hand, many researchers commend the AI Act for its potential to exert a Brussels effect (Floridi, 2021; Hacker, 2023; Wörsdörfer, 2024). Therefore, the AI Act could establish global AI standards by compelling other governments to adopt similar legislative measures or by necessitating multinational companies to comply with European laws worldwide (Hacker, 2023). In terms of protecting young individuals, this implies that legislation with comparable standards could emerge beyond Europe, thus regulating or even prohibiting hazardous AI systems. Consequently, technologies such as deepfakes could be recognized internationally at least as a risk to vulnerable groups. On the other hand, significant gaps persist in the legislation which are already acknowledged by the European Parliament but lack appropriate remedies. In fact, Wörsdörfer (2024) meticulously analyses the strengths and weaknesses of the EU AI law. His research particularly identifies three weaknesses of the AI law that could jeopardise the protection of young individuals, namely the bias of algorithms, the neglect of human rights and the lack of legal remedies. Regarding the latter, Wörsdörfer (ibid) criticises the lack of opportunities for individuals to lodge complaints with market surveillance authorities against the negative impact of AI systems. As a result, vulnerable groups, such as minors, are not provided with sufficient mechanisms or instruments by the AI Act to assert their protection needs.

From this point, substantial literature questions the usefulness of deepfakes for the general public, including whether their disadvantages outweigh their benefits. If this holds true, their risk categorisation should be reconsidered to potentially elevate their status from a transparency risk to a high or unacceptable risk. Indeed, numerous studies underscore the significant drawbacks of deepfakes. De Rancourt-Raymond and Smaili (2023) argue that deepfakes can cause substantial financial losses due to fraud or personality theft among business professionals and companies. Additionally, this type of fraud can affect private individuals through blackmail, impersonation and defamation, leading to severe personal and professional repercussions (de Ruiter, 2021). However, considering deepfakes solely as harmful would be a simplistic and one-sided view. In fact, Mustak et al. (2023) demonstrate that companies can leverage deepfakes profitably, for instance in marketing campaigns and cost-saving measures. Furthermore, they suggest that companies can employ AI influencers to highlight both the opportunities and dangers of deepfakes, thus contributing to the protection of minors through education and awareness than outright prohibition. Therefore, the role of companies in the protection of young people is becoming increasingly important, particularly those that develop or permit deepfake technologies on social media. For instance, there are now detection methods based on algorithms and

neural networks capable of identifying AI-generated media content (Martin-Rodriguez, Garcia-Mojon and Fernandez-Barciela, 2023). Although market-ready recognition tools are still in their early stages, the EU AI law could already promote the mandatory implementation of such recognition tools on social media. This presents significant opportunities not only to reclassify deepfakes into higher risk categories but also to establish oversight bodies that can enhance the protection of minors. (916 words)

References

Bradford, A., 2019. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.

de Rancourt-Raymond, A. and Smaili, N. (2023). The unethical use of deepfakes. *Journal of Financial Crime*, 30(4), pp.1066–1077. Available at: https://doi.org/10.1108/JFC-04-2022-0090.

de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy & Technology*, 34, pp.1311–1332. Available at: https://doi.org/10.1007/s13347-021-00459-2.

Ebers, M., Hoch, V.R.S., Rosenkranz, F., Ruschemeier, H. and Steinrötter, B. (2021). The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J*, 4(4), pp.589–603. Available at: https://doi.org/10.3390/j4040043.

European Parliament, 2024. *Artificial Intelligence Act: MEPs adopt landmark law*. [press release] 13 March 2024. Available at: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law [Accessed 09 June 2024].

Floridi, L. (2021). The European Legislation on AI: a Brief Analysis of its Philosophical Approach. *Philosophy & Technology*, 34, pp.215–222. Available at: https://doi.org/10.1007/s13347-021-00460-9.

Hacker, P. (2021). A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, Innovation and Technology*, 13(2), pp.257–301. Available at: https://doi.org/10.1080/17579961.2021.1977219.

Hacker, P. (2023). The European AI liability directives – Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, 51, pp.1–42. Available at: https://doi.org/10.1016/j.clsr.2023.105871.

Madiega, T., 2024. *Artificial Intelligence Act*. [pdf] European Union. Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf [Accessed 09 June 2024].

Martin-Rodriguez, F., Garcia-Mojon, R. and Fernandez-Barciela, M. (2023). Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks. *Sensors*, 23(22), pp.1–15. Available at: https://doi.org/10.3390/s23229037.

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. and Dwivedi, Y.K. (2023). Deepfakes: Deceptions, Mitigations, and Opportunities. *Journal of Business Research*, 154, pp.1–15. Available at: https://doi.org/10.1016/j.jbusres.2022.113368.

Wörsdörfer, M. (2023). Mitigating the adverse effects of AI with the European Union's artificial intelligence act: Hype or hope? *Global Business and Organizational Excellence*, 43(3), pp.106–126. Available at: https://doi.org/10.1002/joe.22238.

Keywords

Brussels effect: The adaptation of legal norms, regulatory measures and standards of the European Union in countries that are not members of the EU and are not legally bound to the EU law (Bradford, 2019). This influence often occurs because EU laws are so impactful that non-EU governments and multinational companies find themselves forced to comply.

Deepfake: Artificially generated media content, usually videos, images or audio recordings, which are produced using artificial intelligence (AI) or algorithms to convincingly imitate the appearance, voices, facial expressions or gestures of real people. They are often difficult to distinguish from real content and are increasingly used for disinformation.

European Union Artificial Intelligence Act (EU AI Act): The world's first AI law aimed at regulating artificial intelligence (AI) by legal means. The EU AI Act was ratified in March 2024 and categorises AI systems into four risk levels. AI systems that violate fundamental EU principles are categorised as unacceptable risks and are therefore banned (Madiega, 2024).

Transparency risk: The second risk level of the EU AI Act, which includes deepfakes. AI systems categorised under the transparency risk only have to fulfil the obligation that they are clearly recognisable to third parties.