AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective

Milan Tahraoui, Research Fellow at HWR Berlin, Ph.D. Candidate at Paris 1 Panthéon-Sorbonne University and FU Berlin, and Julia Aigner, Research Fellow at HWR Berlin

We argue that, given the legal and institutional challenges that exist at the empirical level, the regulation of AI in the context of judicial proceedings is best conceived from a socio-technical perspective, in order to ensure that the introduction of AI-based tools for deepfake detection and the evaluation of digital evidence can both adequately serve the economy of legal proceedings without disadvantaging parties to legal proceedings and, more generally, without widening technological gaps at the level of society as a whole (European Parliament, 2023, pp. 51-52).

Regulation of deepfake detection in the judicial context must—due to the high sensitivity involved in such context and despite the technical nature of the field of forensics—reflect appropriate legal and ethical requirements for the specific capabilities and limitations of the AI-based tools involved. However, it must also take into account the different actors involved and the institutional settings associated with their deployment. Those actors include judges, prosecutors, defence lawyers, experts, affected persons and, more broadly, oversight bodies, security agencies, civil society organisations, and the public. Considering the institutional context in which AI tools are used does not require adapting the applicable rules to the specific context of use. Yet, this can crucially contribute to the concretisation of what currently constitutes abstract and general requirements under the future EU regulation for AI systems in view to push for an AI regulation that reflect practical, technological and human considerations at stake in a sensitive field of AI application.

One of the most important but challenging objectives in relation to the use of AI-based deepfake detection tools before tribunals is the need to develop differentiated ways to serve the emerging AI principles of transparency and accountability.

Consequently, what is needed for regulating AI systems and developing best practices is the anticipation as far as possible of the concrete risks that may arise in the sensitive contexts in which AI-based tools are used for assessing digital evidence, in order to ensure a more effective compliance with the requirements of the rule of law. A social and an institutional analysis can be combined with a technical understanding of AI tools can serve the overall aim to concretise the legal requirements applicable to high-risk AI systems, which are actually so qualified because of the general potentially of risks arising from their context of use.

In this post, we explore the significance of AI-based detection in the judicial context. In the first section, we begin by defining deepfakes and subsequently provide insights into some of the challenges associated with their detection in the field of digital forensics. In

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

doing so, we address the challenges surrounding the admissibility and reliability of evidence in court, and highlight the practical difficulties of distinguishing genuine content from manipulated media. In section two, we look at the regulatory landscape, discussing the future EU legal framework and highlighting the need for practical guidance. Section three focuses on the specific challenges of deepfakes in the judicial arena, considering the roles of different stakeholders in the legal process. Under section four, we argue for a socio-technical perspective on deepfake detection, emphasising the integration of human expertise with technological advances and offer recommendations for a balanced approach.

1. AI-based detection: What is at stake for the independence and impartiality of the judiciary?

The term deepfakes generally refers to visual content that is artificially created, manipulated or distorted by using artificial intelligence tools to alter or replace a person or selected attributes of that person in the content. Deepfakes encompass both visual (i.e. images and videos) and auditory (i.e. sounds and noises) content (Gregory & Cizek, 2023; Witness & MIT, 2023). The European Parliament recently proposed an amended definition for the future AI Regulation: "'deep fake' means manipulated or synthetic audio, image or video content that would falsely appear to be authentic or truthful, and which features depictions of persons appearing to say or do things they did not say or do, produced using AI techniques, including machine learning and deep learning"(European Parliament, 2023, p. 124).

Despite attempts to define deepfakes in various jurisdictions, there is no consensus on what exactly a deepfake is. One of the reasons for this is that deepfakes are a sub-category of synthetic media (Partnership on AI, 2023, p. 3). In practice, it is difficult to distinguish between digital content that has been slightly modified with the help of AI tools, e.g. through filters, and synthetic media and deepfakes, which are generally considered to be more transformed or artificial. Media policies imposed by major social networks illustrate this difficulty. Several social networks have officially banned deepfakes on their platforms, while allowing AI-modified content. The widespread use of filters on social networks, which makes it easy to use deepfake technologies when creating content, makes it difficult to draw a clear distinction. All of these definitional subtleties will come into play when assessing specific digital content as potential digital evidence to prove the preparation or commission of international crimes or offences under some national jurisdictions, deepfake-based criminal offences or general criminal offences. However, one common definitional element specific to deepfakes is the use of

(AI) technology to confuse, mislead or manipulate an audience based on their perception of reality.

As the core non-technical definitional element characterising deepfakes, the primary intent behind deepfakes varies. While there are legitimate uses, such as in film

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

production or in voice synthesis for virtual assistants, satirical purposes, and political or social advocacy, deepfakes are also generated and/or circulated for malicious purposes like sexual harassment, defamation, fraud, disinformation, and manipulative campaigns (Witness & MIT, 2023; Tahraoui et al., 2023). In the context of judicial proceedings, parties may, wittingly or unwittingly, introduce deepfakes as evidence, which in turn can influence the ultimate outcome of a case.

In order to counteract the dangers, deepfake detectors are used to recognise manipulated media. With AI-based deepfake detection, law enforcement authorities and courts can make probability predictions regarding the authenticity of submitted image and video data when used optimally. The detection tools use advanced algorithms to identify anomalies in the visual characteristics of a video or image that could indicate manipulation. For example, they analyse a person's eye movements, lighting, or other subtle details to determine whether the footage is authentic or not (Kraetzer et al., 2022; Ramachandran et al., 2021; Yu et al., 2021).

In recent years, there has been an explosion in the use of digital evidence in legal proceedings. While in the past it was rather common in international criminal proceedings, due to the great difficulties of conducting investigations on the ground, it is now increasingly becoming one of the most common sources of evidence in both civil and criminal proceedings at the national level. The rise of deepfake technologies is already structurally bringing challenges for using and assessing digital evidence in court. Even before deepfakes became widespread, assessing the authenticity and probative value of digital evidence was already a complex task in practice. As with other consequences of the digitalisation of society and social interactions, deepfakes in the judicial context do not create entirely new challenges and problems, but rather amplify existing ones. The rapid progression and quality of deepfake technologies make it difficult for the unaided expert human eye to distinguish between genuine and manipulated media. There are also concerns about biases within AI-based detection tools and the potential risks of consequently convicting individuals based on evidence that is in fact deepfakes (Delfino, 2023). As deepfake creation tools become more advanced, the cat-and-mouse game between creation and detection intensifies (Masood et al., 2023; Passos et al., 2022). This evolution presents multifaceted challenges:

Authenticity and probative value of digital evidence: Law enforcement authorities and courts face the complex task of detecting deepfakes to obtain or secure digital evidence of satisfactory quality, especially given the absence of standardised or formalised methods for digital forensics (Lyle et al., 2022). Digital forensic tools can face several risks of errors when applied to digital content, which can be manipulated in various ways, even without being transformed by AI. For instance, metadata is stored in standardized formats and can be easily manipulated, without leaving any apparent traces. Consequently, metadata analysis as evidence may not be useful in certain situations (ENFSI, 2021, p. 16). Inaccuracy is another significant risk when assessing digital

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

evidence, as forensic tools may inaccurately report information, including non-existent artifacts or clustering unrelated items. Alteration of data, such as updating a date/time stamp to the current date, is yet another risk that forensic tools can introduce (Lyle et al., 2022, p. 44).

Deepfake Defence: The phenomenon of "deepfake defence" covers the use of deepfakes to raise objections and arguments relating to the credibility of digital audio and image files, as a potentially legitimate defence strategy within legal proceedings. This defence strategy is expected to see a rise in courtroom scenarios in the near future and can challenge the pursuit of truth within the legal system. The deepfake defence can be seen as a variation of the phenomenon known as the "liar's dividend" (Chesney & Citron, 2019). The "liar's dividend" describes advantages that deceitful individuals gain amidst the widespread uncertainty surrounding the authenticity/credibility of online information. The deepfake defence primarily focuses on the use of deepfakes as a legal strategy, whereas the "liar's dividend" encompasses the broader benefits deceitful individuals acquire from the prevailing skepticism towards online content.

Legal Implications of Deepfake Detection: From a fundamental rights and democratic perspective, the concept of deepfake detection is legally contentious. It delves into using AI technology to discern truth from falsehood. This task is of primordial significance in the judicial context, especially in the continental European legal tradition, where the traditional aim is to establish the truth through a legal process. The role of courts in determining, as far as possible, the truth in legal proceedings is in itself a democratic issue. That said, courts cannot be considered in isolation from the rest of society. Indeed, much of the digital evidence in courts originates from social media or other widely accessible online sources. Concretely, one typical context in which deepfakes will intervene in a judicial context is the use of a deepfake-content that was originally shared online. The democratic dimension of assessing the authenticity and probative value of digital evidence used to establish truth in a judicial context is therefore intertwined with the broader democratic problematic of regulating how online information is circulated and used. A collaborative project on satire and deepfakes between the MIT and the NGO Witness entitled *Just joking!* asked the following relevant questions: "What is the role of government in regulating how different forms of media circulate online? Should legislation dictate protocols for categorizing and vetting AI-enabled media specifically, and/or set standards for accountability? [...] How can democratic societies avoid the weaponization of "deepfake panic" and the "liar's dividend" as well as governmental or commercial policies that overly restrict satire as well as truthful journalism and fact-finding?" (Witness & MIT, 2023). Indeed deepfake detection in courts raises pressing questions as to the role of the government in determining what is factual or fiction in society, a distinction blurred further by deepfakes (Porter & Tan, 2023; van der Sloot & Wagensveld, 2022). These questions are even more sensitive in the case of international crimes, where the structural obstacles to establish truth or falsity are very high.

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

Digital forensics in the deepfake age: AI promises significant advancements in digital forensics, such as the ability to quickly and accurately analyse vast amounts of data. However, it is essential to strike a balance between human expertise and AI to ensure that the justice system remains fair and impartial, even if simply including a human in the loop cannot solve all issues. On the technical side, an important characteristic of digital forensics in general is the lack of uniform, standardised methods for assessing the authenticity and reliability of digital evidence (Lyle et al., 2022). Digital forensic methods are derived among others from computer sciences, which means that methods are developed, tested, and used in an iterative process based on the existing state of the art in relevant scientific fields for solving specific technical questions.

This situation is further complicated by the fact that evidentiary rules and standards in the judicial context are mostly broad, being national or international level, often based on general principles of law and subject to a degree of discretion and reasonableness on the part of judges (Delfino, 2023; Fallah, 2020).

The difficult task of identifying deepfakes while assessing digital evidence in the judicial context may also extend the problem of "partisan scientific evidence" far beyond where it was originally most common, namely in highly technical judicial contexts and cases (Sulyok, 2021). In most national legal contexts, it is highly questionable whether the necessary institutional and human skills and knowledge exist to deal with the problem of the ever-increasing possibilities of manipulating and distorting digital content while making it appear to be true. However, it can be argued that the deepfake phenomenon is an invitation to address these challenges in the medium to long term.

2. The emerging legal regulation of deepfake detection and generation in need of further practical guidance

As the reliance on digital data as evidence continues to grow, there is a need for specialised procedures in acquiring and handling digital evidence to ensure their admissibility in court and evaluate their probative value.

The legal landscape currently lacks specific rules addressing the novel challenges introduced by deepfakes. This creates an environment of uncertainty for both the creators and detectors of deepfakes. Policymakers and legal scholars must grapple with these emerging concerns to draft comprehensive guidelines. Courts mostly resort to general procedural laws, relying heavily on judicial discretion, whereas AI-based deepfake detection tools are not completely explainable for instance. This approach, while rooted in established legal traditions, may not be sufficiently robust to address the intricacies of deepfake detection and generation. It might be well advised to introduce irritation regarding how deepfake detection tools can be used in courts, or what literature called a "go-slow-and-strict approach to the rules of authenticity" (Delfino, 2023, p. 316).

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

In the absence of top-down regulatory guidelines regarding the use of AI systems in their criminal investigative work, law-enforcement authorities can and do take the initiative to develop best practices (Council of Europe, 2020, p. 17). Europol has argued, that given their expertise on digital evidence, law-enforcement authorities can be trusted to develop standards regarding deepfake detection in criminal matters (Europol, 2022, pp. 15-16). However, the very fact that law-enforcement authorities are directly involved in criminal proceedings suffices to highlight the danger of accepting internally developed standards, as a general rule for the evaluation of digital evidence in a civil or criminal trial (Council of Europe, 2020, p. 17), following a so-called "we know our job" approach. Nevertheless, these standards can be legitimately utilised in the internal work of law-enforcement authorities. There is consequently a need for legal regulation in the judicial context, given the concrete but fundamental importance of the authenticity and probative value of digital evidence for rendering an independent and impartial justice. The absence of a comprehensive legal framework tailored specifically for deepfake detection in the evidentiary context is problematic. There are nonetheless some developments at the EU level with the future AI Regulation of the Union, which will cover deepfake detection and generation for evidence analysis in several sensitive use-cases will be covered in its scope of application, including when used in law-enforcement and judicial contexts.

Even when deepfake detection will fall under the scope of application of the future AI Regulation, the legal status of deepfake detection in the EU is still uncertain as negotiations between the European Commission, EU Council, and European Parliament continue. The final version of the AI-Act has not been adopted yet. While deepfake detection has been removed from the list of high-risk AI systems contains in Annex III in the amended version of this Regulation by the EU Council, it can still fall under Art. 6 lit. d of this Annex that deals with the use of AI systems for assessing the reliability of evidence in criminal proceedings (European Parliament, 2023, p. 333). The European Parliament suggests that deepfake detection generally falls under the medium risk category but may be partially reclassified as high risk if used in the forensic work of law enforcement agencies. Similarly, the high-risk qualification will be applicable to AI systems used by the judiciary, as this transpires in the amended Recital 40: "AI systems intended for the administration of justice and democratic processes should be classified as high-risk, considering their potentially significant impact on democracy, rule of law, individual freedoms as well as the right to an effective remedy and to a fair trial. In particular, to address the risks of potential biases, errors and opacity, it is appropriate to qualify as high-risk AI systems intended to be used by a judicial authority [...] in researching and interpreting facts and the law and in applying the law to a concrete set of facts or used in a similar way in alternative dispute resolution" (European Parliament, 2023, pp. 54-55). One crucial aspect of this context is the handling of evidence in legal proceedings and the various checks and balances that courts must apply. As a consequence of their qualification as high-risk systems, AI systems used for assessing digital evidence will have to comply with the security, transparency, and oversight obligations set forth in Chapter II of the AI Regulation, including having a risk

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

management system, using high-quality data models, avoiding bias, and providing functions for human oversight and interpretation of results (Roos & Weitz, 2021).

In addition, the future AI regulation will require deepfakes to be labeled as such when they are In addition, the future AI regulation will require deepfakes to be labeled as such when they are circulated, unless when they are used for law-enforcement purposes or fall under a fundamental freedom exception (European Commission, 2021, p. 69). Given the risks to personal and societal interests, it is important to consider if transparency obligations alone are sufficient to tackle challenges relating to the assessment of digital evidence in courts, even in the optimistic scenario in which users and platforms will comply with labeling obligations. There are unanswered questions, such as whether this obligation applies to various types of deepfakes, ranging from smiley-faced suns to fully AI-generated deepfake persons. Furthermore, it is unclear who the addressees for receiving this information are: is it the general public, the person depicted, or the platform? Furthermore, the method and format for disclosing the information—whether it should simply state that it is a deepfake or specify which part of the video, image, or audio has been manipulated—remain uncertain (van der Sloot & Wagensveld, 2022). The exponential generation and use of synthetic media and the increasing use of AI tools to modify online content will further complicate this.

The future AI Regulation will also impose transparency obligations and requirements for deepfake detection tools, but there are still many uncertainties regarding practical implications stemming from their use in the judiciary context. Specific regulations for AI-based systems in court proceedings will have to be in place at both the European and national levels.

Despite the expected developments of future guidelines tailored to sensitive areas like the police and judiciary, implementing key obligations in relation to digital evidence and AI systems can be challenging. The EU Commission is expected to offer clearer guidance on labellng deepfakes when they are shared online, specifically concerning Article 52(3) of the future AI Regulation (European Parliament, 2023). These guidelines aim to resolve implementation issues related to deepfakes, particularly in terms of presenting evidence in legal settings. However, there may be practical and legal obstacles when it comes to implementing EU law in national contexts. It is important to consider the overall conversation about using digital data as evidence in court when developing and implementing new procedural laws for deepfake detection. When it comes to the admissibility of evidence that violates the GDPR, for example, such evidence has been deemed admissible in multiple criminal courts (Haag, 2022). Recent cases such as

EncroChat and SkyECC indicate a growing willingness from investigative authorities to employ special investigative techniques in criminal cases, involving for instance sophisticated hacking techniques and international cooperation. These developments

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

may create opposition to the increased standardisation or proceduralisation of the assessment of digital evidence usage in court proceedings.

Practical challenges persist. The reliability and potential biases of deepfake detectors, coupled with the inherent opacity of many AI models, raise questions about their admissibility in courts. The need for transparency and the ability to scrutinize evidence are paramount in legal contexts, and AI models must meet these standards to be considered reliable.

3. Deepfake detection and generation in the judicial context: challenges for assessing admissibility and probative value of digital evidence

The use of AI-based systems, particularly in criminal enforcement and legal proceedings, brings about significant risks to basic human rights (Brooks et al., 2021; van der Sloot & Wagensveld, 2022). The use of AI-based deepfake detection can for example target the right to a fair trial and the equality of arms in criminal proceedings (Delfino, 2023). Therefore, the importance of thinking through and adapting procedural requirements to serve substantive justice goals must be stressed.

The complexities of AI-based systems, often likened to a 'black-box', present challenges for their use in the judicial contexts. While AI tools offer a range of technical solutions for digital forensics, they cannot supplant the role of human expertise in court (European Parliament, 2023, pp. 54-55). Experts serve as crucial intermediaries, deciphering and contextualizing AI outputs, ensuring that the evidence presented is both reliable and admissible (European Commission, 2022). But in practice, can we expect courts to have sufficient human expertise to counterbalance the opacity, the risk of error, and to simply make sense of deepfake detection tools? Similarly, how can the right to equality of arms in trials be served in the context of the introduction of AI tools to deal with one of the concrete cornerstones of justice: evidence to establish of facts? In addition, how can the administration of justice be better served, given the risks that for instance the deepfake defence may lead to an increase in the time and costs associated with proceedings, without compromising the rights of the parties to the trial and the public interest (Haag, 2022)?

In any case, deepfake detection tools cannot be summed up simply in terms of their technical capabilities and performance: they are not simply products to be used in court, although they must include technical guarantees as to their consistency, accuracy and reliability over time (Nalepa et al., 2019). The conditions for generating trust mandate rigorous scrutiny, ensuring that AI-based tools meet the stringent standards required for evidence in judicial proceedings (Laux, 2023). In that context, the relationship between

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

AI-based tools and human needs to be at the centre of a legal perspective over the regulation of deepfake detection in courts.

The accessibility to AI-based detection tools poses another challenge. Obtaining and integrating these tools requires substantial financial resources, which may be more feasible for public institutions but often limited for individuals. As a result, disparities can arise, leading to an unbalanced representation in judicial proceedings. This, in turn, can impact the credibility and dependability of presented evidence, as well as its evaluation by the judiciary (European Parliament, 2023, p. 52).

Incorporating deepfake detection tools into legal proceedings necessitates careful consideration, especially regarding procedural requirements. The principles of justice, fairness, and equity must remain central. It is imperative to strike a balance between embracing technological advancements and upholding the foundational tenets of the legal system, but operationalising this goal requires further consideration of how deepfake detection tools will be used in practice, by taking into account the social context of their use.

4. Embedding interactions between human expertise and technological tools for deep fake detection in legal proceedings from a socio-technical perspective

For our purpose using a socio-technical perspective means framing the regulation of deepfake detection in the judicial context by focusing on how human interaction with technology is shaped by the broader societal context (Schwartz et al., 2022). In this context, key challenges for assessing digital evidence include:

Transparency: The transparency of AI tools, especially regarding deepfakes, is a major concern. As raised in discussions surrounding the European Parliament's draft amendments, numerous questions emerge: How should deepfakes be labeled? Should a distinction be made between a wholly AI-generated individual and a manipulated real person? How detailed should the disclosure be—just a generic "deepfake" label or an in-depth breakdown of the manipulations? These questions underscore the complexities in ensuring transparency in AI outputs.

Human-AI Collaboration: Experts not only validate AI outputs but also serve as crucial mediators, interpreting and contextualising AI results for legal professionals. This connection ensures that while AI offers technical solutions, human experts incorporate ethical, legal, and contextual considerations. However, this interdependence between humans and AI systems can sometimes blur boundaries. In certain cases, the human operator assigned to control an AI system may become integrated into the AI operating system itself, rather than acting merely as an external oversight person (Van Den Meerssche, 2022, pp. 186-187).

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

Understanding the AI Lifecycle: Effective AI regulation requires considering the entire AI lifecycle. This includes everything from development and training to deployment and post-deployment assessments. This comprehensive view ensures that regulations address potential pitfalls at every stage, from biases in training data to challenges in real-world applications.

Accountability and Best Practices: As AI tools wield significant influence in legal proceedings, it is paramount to ensure they adhere to best practices. This includes rigorous testing, continuous validation, and a commitment to ethical considerations as well as respecting the legality principle in criminal matters. Accountability mechanisms must be in place to address potential errors or biases in AI outputs.

Roles of Various Stakeholders: The deployment and utilization of AI tools in legal settings involves multiple stakeholders—from AI developers and forensic experts to judges, lawyers, and even the general public. Each plays a pivotal role, bringing unique perspectives and expertise to the table. Policymakers and regulators, for instance, shape the environment in which these tools operate, ensuring they align with societal values and legal norms. Specifically, advocating a socio-technical approach calls for not putting the responsibility solely on AI developers or AI (end-) users. A socio-legal perspective encourages framing responsibility in the legal context based on the various actors that need to be involved at different levels of the AI life cycle (Krätzer et al., 2023). This means taking into account practical and institutional factors that shape the materialisation and mitigation of risks when abstract regulatory requirements are applied in practice.

AI deployers and administration of justice: We argue that the newly proposed category of AI deployers should in particular be used to further develop this socio-technical perspective. The category of AI deployers has been proposed by an amendment that basically replaces the term of AI users with that of AI deployers in the version of the AI Regulation agreed to by the European Parliament, without deleting mentions of AI users systematically. AI deployers are the persons or entities responsible for deploying high-risk AI systems on the ground, with the specific task of ensuring that AI systems are used in compliance with ethical and legal requirements, while at the same time analysing and formulating the practical needs for the use of AI systems (European Parliament, 2023). AI deployers can be thought of as the hierarchy of an administration. In the judicial context, AI deployers could be judges, the registry or any organ in charge of the administration of justice that will have the power and duty to regulate and/or supervise the use of AI tools in the concrete work of tribunals. AI deployers have a crucial role in ensuring the protection of fundamental rights. According to amendment recital of the European Parliament, they need to complement the responsibilities of the provider by understanding how the system will be used and identifying unforeseen risks. This requires knowledge of the specific context, the potential impact on different groups, and the implementation of appropriate governance structures. To ensure fundamental rights are safeguarded, deployers should conduct a thorough impact assessment before implementing the system (European Parliament, 2023, pp. 69-70).

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

Conclusion

In this post, we argued that the regulation of AI-based deepfake detection tools in the context of legal proceedings requires a socio-technical perspective. This is necessary in order to clearly distinguish the needs arising from the use of AI tools, such as those used for deepfake detection in legal proceedings, from those related to AI systems used in other contexts, such as for self-driving cars. Specifically, this means moving away from a regulatory model that considers the technical development and sale of AI tools in isolation from their use in the field. To this end, analysing the empirical conditions of the use of AI-based deepfake detection tools, taking into account the institutional and human factors that influence their concrete use and the risks arising from their deployment in courts is essential.

Traditionally, algorithmic tools are regulated in a compartmentalised way. However, the complexity of developing and operating AI systems clearly demonstrates the importance of regulating AI tools considering their distinct "life cycles" and the overlapping roles of different stakeholders involved in these life cycles. From development to deployment and end use, it is crucial to recognize the risks associated with placing excessive emphasis on requirements for AI developers or end users alone, without considering the necessary institutional integration of AI systems and the intermediary responsibilities that exist between development and final usage. Instead, a comprehensive regulatory framework should address the intermediary steps and responsibilities, ensuring the technology's ethical and efficient integration into our legal systems. Ultimately, as AI systems are increasingly integrating our judicial processes, it is imperative to strike a balance. While AI offers immense potential benefits, its application must be approached with caution, understanding, and an unwavering commitment to upholding the principles of justice.

Bibliography:

- Brooks, T., G., P., Heatley, J., J., J., Kim, S., M., S., Parks, S., Reardon, M., Rohrbacher, H., Sahin, B., S., J., T., O., & V., R. (2021). Increasing Threat of DeepFake Identities. U.S. Department of Homeland Security.https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf
- Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 175–1820.
- Council of Europe, Parliamentary Assembly (2020). Justice by algorithm – the role of artificial intelligence in policing and criminal justice systems (Doc. 15156). https://pace.coe.int/pdf/49efbf908c3b5d9c5251903aecf86a67a1d15a873e48a18c6d21d3e560eb8bab/doc.%2015156.pdf
- Delfino, R. (2023). Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery. Hastings Law Journal, 74(2), 293–350. https://doi.org/10.2139/ssrn.4032094
- ENFSI (2021). Best Practice Manual for Digital Image Authentication ENFSI-BPM-DI-03. Best Practice Manual, 2021(01).
- European Commission, Directorate General for Communications Networks, Content and Technology (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. Publications Office. https://data.europa.eu/doi/10.2759/002360
- European Parliament (2023, Mai 11). European Parliament, Draft Compromise Amendments on the Draft Report, Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD))', KMB/DA/AS, 11 May 2023. https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.

- Europol (2022). Facing reality? – Law enforcement and the challenge of deepfakes: An observatory report from the Europol innovation lab. Publications Office. https://data.europa.eu/doi/10.2813/08370
- Fallah, S. M. (2020). The Admissibility of Unlawfully Obtained Evidence before International Courts and Tribunals. The Law & Practice of International Courts and Tribunals, 19(2), 147–176. https://doi.org/10.1163/15718034-12341420
- Haag (2022). Teil 29.5 KI im Strafrecht und Strafprozessrecht. In Handbuch Multimedia-Recht (58. Aufl.).
- Witness & MIT (2023). Deepfakes, Satire, and the Politics of Synthetic Media.. https://cocreationstudio.mit.edu/just-joking/
- Kraetzer, C., Siegel, D., Seidlitz, S., & Dittmann, J. (2022). Process-Driven Modelling of Media Forensic Investigations-Considerations on the Example of DeepFake Detection. Sensors, 22(9), 31–37. https://doi.org/10.3390/s22093137
- Krätzer, C., Siegel, D., Seidlitz, S., Dittmann, J., Siegel, D., Seidlitz, S., & Dittmann, J. (2023). Human-in-control and quality assurance aspects for a benchmarking framework for DeepFake detection models. Electronic Imaging, 35, 379-1–379-6. https://doi.org/10.2352/EI.2023.35.4.MWSF-379
- Laux, J. (2023). Institutionalised Distrust and Human Oversight of Artificial Intelligence: Toward a Democratic Design of AI Governance under the European Union AI Act. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4377481
- Lyle, J. R., Guttman, B., Butler, J. M., Sauerwein, K., Reed, C., & Lloyd, C. E. (2022). Digital investigation techniques: A NIST scientific foundation review (NIST IR 8354; S. NIST IR 8354). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.IR.8354
- Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, 53(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z
- Nalepa, G. J., Araszkiewicz, M., Nowaczyk, S., & Bobek, S. (2019). Building Trust to AI Systems Through Explainability. Technical and legal perspectives. XAILA@JURIX. https://www.semanticscholar.org/paper/Building-Trust-to-AI-Systems-Through-Technical-and-Nalepa-Araszkiewicz/7fd5056271090809d187f0b70cb39689ed848390
- NBC News (2023, März 3). AI-generated “synthetic media” is starting to permeate the internet. NBC News. https://www.nbcnews.com/tech/tech-news/ai-generated-synthetic-media-future-content-rcna72958
- Partnership on AI, PAI’s Responsible Practices for Synthetic Media: A Framework for Collective Action. (2023, Februar 27). Partnership on AI. https://partnershiponai.org/resource/pais-responsible-practices-for-synthetic-media-a-framework-for-collective-action/
- Passos, L. A., Jodas, D., da Costa, K. A. P., Júnior, L. A. S., Colombo, D., & Papa, J. P. (2022). A Review of Deep Learning-based Approaches for Deepfake Content Detection. https://doi.org/10.48550/ARXIV.2202.06095
- Porter, T., & Tan, N. (2023). Introduction: Democracy and digitization. Globalizations, 20(1), 132–136. https://doi.org/10.1080/14747731.2023.2156033
- Ramachandran, S., Nadimpalli, A. V., & Rattani, A. (2021). An Experimental Evaluation on Deepfake Detection using Deep Face Recognition. 2021 International Carnahan Conference on Security Technology (ICCST), 1–6. https://doi.org/10.1109/ICCST49569.2021.9717407
- Roos, P., & Weitz, C. A. (2021). Hochrisiko-KI-Systeme im Kommissionsentwurf für eine KI-Verordnung IT- und produktsicherheitsrechtliche Pflichten von Anbietern, Einführern, Händlern und Nutzern. MMR Zeitschrift für IT-Recht und Recht der Digitalisierung, 2021, 844–851.
- Schwartz, R., Vassilev, A., Greene, K. K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. NIST. https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence
- Sulyok, K. (2021). Science, Legitimacy and the Judicial Function—A Need for More Intrusive Standards of Review (SSRN Scholarly Paper 3895130). https://doi.org/10.2139/ssrn.3895130
- Tahraoui, M., Krätzer, C., Dittmann, J., & Aden, H. (2023). Defending informational sovereignty by detecting deepfakes. WEIZENBAUM JOURNAL OF THE DIGITAL SOCIETY, 3(2). https://doi.org/10.34669
- Van Den Meerssche, D. (2022). Virtual Borders: International Law and the Elusive Inequalities of Algorithmic Association. European Journal of International Law, 33(1), 171–204 https://doi.org/10.1093/ejil/chac007
- van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: Regulatory challenges for the synthetic society. Computer Law & Security Review, 46. https://doi.org/10.1016/j.clsr.2022.105716
- Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A Survey on Deepfake Video Detection. IET Biometrics, 10(6), 607–624. https://doi.org/10.1049/bme2.12031

Cite as: Milan Tahraoui and Julia Aigner , AI-based Deepfake Detection in Judicial Proceedings: A Socio-Technical Perspective, *Limablog*, May/2024, p. 1-12, available at: <https://limablog.org/ai-based-deepfake-detection-in-judicial-proceedings-a-socio-technical-perspective/>.