Closing the Gaps: Proposals for Enhancing the EU Artificial Intelligence Act

Valerie Megan[1,*,†], Youki Kadobayashi[1,†]

[1]Division of Information Science, Nara Institute of Science and Technology (NAIST), Ikoma, Nara, Japan

Abstract

The Artificial Intelligence Act (AIA) is a pioneering legislative framework proposed by the European Union to regulate artificial intelligence (AI) technologies, particularly high-risk AI systems. Although this Act sets a model for global legislation, our findings reveal that there are vital vulnerabilities. These critical areas include vague definitions and potential exemptions that might lead to loopholes. Our main contribution is in the analysis of AIA for both layman and researchers for possible solutions and insights, aiming to provide an overview of how AI regulation is being dealt with beyond the EU, enriching the analysis that leads to better legislation on the AIA. Another issue raised is regarding virtual influencers, how the AIA should treat them, and what dangers they might entail. We aim to make the AIA more efficient by preventing misuse, ensuring fundamental rights, and not restraining innovation. Our recommendations offer detailed concrete strategies that would ensure the Act remains effective and responsive to changes in future AI advancements.

Keywords

artificial intelligence (AI), European Union (EU), Artificial Intelligence Act (AIA), regulation

1. Introduction

The application of science-based technologies has significantly improved our lives, more so recently, with Artificial Intelligence (AI) altering personal and professional aspects. Despite its potential, such as in medical imaging and autonomous vehicles, careful handling is crucial to avoid unintended consequences, such as bias in recruitment tools or faulty facial recognition systems.

Proposed on April 21, 2021, by the European Commission, the Artificial Intelligence Act (hereinafter: AIA) is the first in-depth law on the matter, aiming to strengthen existing rights and promote trustworthy AI through setting specific requirements and obligations. According to a press release of the European Parliament dated March 13, 2024, the AIA was overwhelmingly approved. Whilst awaiting a final legal-linguistic review, it is expected to be formally adopted before the current parliamentary term concludes. Following its publication in the official journal, it will take effect within 20 days. Full implementation will be phased over a 24-month interval, with stricter regulations for high-risk systems coming into force after 36 months [1].

Our analysis identified some vulnerabilities that could be exploited by malicious actors. This position paper further elaborates on such vulnerabilities to set a stage for further discussion and mitigation strategies. Therefore, our main contributions are analyzing AIA from the perspectives of both laypersons and researchers, potential solutions and insights, global comparison of regulations concerning AI, and raising awareness of threats related to virtual influencers (VIs). This paper is intended to inform legislators and researchers, thus encouraging cooperation that can help make the AIA more effective and resilient.

As a position paper, we outline current AIA challenges and propose research areas for the AI community:

AIEB 2024: Workshop on Implementing AI Ethics through a Behavioural Lens | co-located with ECAI 2024, Santiago de Compostela, Spain

*Corresponding author.

†These authors contributed equally.

valerie.megan.vp9@naist.ac.jp (V. Megan); youki-k@is.naist.jp (Y. Kadobayashi)

0009-0002-7198-5049 (V. Megan); 0000-0001-9630-4831 (Y. Kadobayashi)

1. **AI Detection and Compliance:** How can AI detection methods be advanced to reliably identify AI-generated content across various media types, ensuring consistent compliance with regulations?
2. **Virtual Influencers and AI-Generated Personas:** What are the effective methods for detecting whether a user is a natural person or an AI-generated persona? Are the current detection algorithms developed by researchers capable of reliably identifying AI-generated personas? How human-like are current AI-generated personas, and what advancements are needed to further improve their quality and realism?

By addressing these challenges and implementing proposed solutions, the AI community can contribute to evolving AI regulations, ensuring they are robust, adaptive, and capable of safeguarding against emerging threats while promoting innovation.

2. Methodology

Our research methodology involved several steps:

1. **Literature Review:** We reviewed existing literature on AI regulation, focusing on scholarly articles, legal analyses, and policy papers related to the AIA and other global AI regulatory frameworks.
2. **Comparative Legal Analysis:** We compared the AIA against established AI regulations from other jurisdictions, such as the NIST AI Risk Management Framework (USA), the Model AI Governance Framework (Singapore), and the G7 framework.
3. **Content Analysis:** We analyzed the AIA text using content analysis techniques to identify ambiguities, potential loopholes, and areas lacking clarity. This involved developing a coding framework, segmenting the text, and thematic coding.
4. **Expert Consultation:** We engaged with experts in defense alliances and cybersecurity to validate our findings regarding the military section.
5. **Case Studies:** We incorporated case studies of recent AI-related incidents, such as the misuse of deepfake technologies involving Taylor Swift and military AI deployments by Ukraine.

3. EU-Defense Alliance Collaboration in Military AI

In late 2023, Ukraine's deployment of the AI-powered Saker Scout drones on Russian territory highlighted accountability issues in the use of advanced military technologies. These drones, designed for enhanced target engagement and resistance to signal jamming, autonomously process data to improve decision-making in disrupted environments [2, 3]. Concurrently, DARPA's LongShot program developed similar autonomous capabilities in unmanned aerial vehicles, avoiding regulatory oversight and emphasizing the need for robust frameworks to manage AI in military contexts [4, 5].

The NATO Autonomy Implementation Plan aims to adopt autonomous systems for defense, emphasizing norms, values, and international law [6]. However, dual-use civilian-military technologies, like drones and GPS, pose regulatory challenges and security threats due to their adaptable nature. These technologies, often developed by non-state providers, undergo extensive safety and accountability reviews during peacetime but exempted in wartime.

Acknowledging these complexities, exempting military AI from the scope of the AIA was a strategic decision by the EU to mirror the different roles that military AI has been allowed to evolve under the facilitative role played by the defense alliance. It was an important strategic decision that highlights a very important policy difference within EU Member States about how to regulate different types of military technologies. Since missiles, tanks, and guns have traditionally been regulated at the national level due to their direct association with national defense, AI introduces new complexities to its dual-use nature and the involvement of civilian technology providers. While dual-use technologies and private

sector involvement aren't new, AI's black-box nature adds unique complexities. Despite the efforts of Explainable AI (XAI), AI systems often remain opaque, exacerbating complexities in handling dual-use technologies. Such a discrepancy calls for an evolved regulatory framework that not only meets the advances being made in AI but also is in line with present regulations for traditional military hardware [7].

Instead of total exemption in the AIA, a collaborative EU-Defense Alliance framework could be devised wherein military AI has undergone stringent testing and certification processes for ethical compliance, operational safety, and legal accountability. This collaborative framework would precisely describe the areas of roles for the EU and the defense alliance: (1) while the defense alliance could be responsible for overseeing the deployment of military AI systems in terms of responsible use and compliance with international law, (2) the development work could be led by the EU. This is fundamental, taking into account that military AI systems are often developed by non-state military providers, hence requiring strict EU control to make sure that such technologies are of a nature corresponding to the high standards required for military application.

In conclusion, as the defense alliance continues to play a crucial role in the strategic use of military AI, the EU must advocate for a shared testing and regulatory framework to align military AI technologies with the highest standards of international law and ethical norms while enhancing the operational capabilities of EU and defense alliance forces. It will ensure ongoing oversight and accountability and continue developing military technology innovations. Adopting this approach will eventually guarantee an all-rounded governance framework dealing effectively with the details and complications of modern military operations while providing an ethically aligned deployment of AI in the defense sectors.

4. Oversight in AI Media

4.1. Personal Activity Exemption

In late January 2024, platform X was flooded with sexually explicit deepfake images of Taylor Swift, accumulating at least 45 million views before mitigation. These images were traced to Microsoft's AI-driven text-to-image generation model. This model uses advanced generative adversarial networks (GANs) to synthesize highly realistic images by training on large datasets of facial features, producing outputs that closely resemble real human appearances.

GANs consist of two competing neural networks: the generator and the discriminator. The generator creates images, while the discriminator distinguishes between real and generated images, improving the generator's output through iterative adversarial training. This understanding is crucial for regulators to identify potential misuses of GAN technology [8].

The incident raises critical questions under the AIA's current legislative framework, specifically regarding the exemption for 'purely personal non-professional activity' in Articles 2(5c) and 3(4) AIA. Zubear Abdi, identified as Zvbear on X, might argue that his use of Microsoft's AI model falls within this category, as it was not for monetization or serving others but created out of boredom, without receiving benefits like money or fame. By this definition, his activity is purely personal, exempting him from AIA compliance.

This situation highlights the necessity for integrating technical insights into AI functionalities to prevent misuse and protect individuals from unauthorized digital representations. Further technical and ethical standards are needed to align AI capabilities with legal frameworks that address both the potential and risks of AI in content creation. Legislators must review and refine the wording of this provision to clearly define the scope and limits of activities that qualify for exemption.

4.2. Content Generation and Marking Obligations

Article 52(1a) AIA mandates marking AI-generated outputs, with exceptions for systems assisting with standard editing, minimal alterations, or lawful purposes. However, ambiguity in the definition

of "content generation" allows some systems to claim exemption by repurposing pre-recorded or human-created content.

Moreover, the effectiveness of current AI output marking techniques is questionable. These techniques are vulnerable to alterations or manipulations, potentially leading to false assertions of compliance. The broad interpretation of exceptions allows for the design of systems that, while technically complying with the letter of the law, may significantly alter content in a manner that arguably defies the spirit of the regulation.

Several algorithms detect deepfakes in visual media, such as FakeCatcher, which uses photoplethysmography to detect synthetic changes by appraising blood flow patterns [9]. PhaseForensics analyzes lip motion frequency through neural networks [10], and TruFor detects AI-generated and manually manipulated images by identifying digital signatures [11].

In audio, vocal tract reconstruction checks sample authenticity against biological reality [12], while speaker verification systems extract biometric characteristics from recordings [13]. Textual deepfake detection includes watermarking with green-listed vocabularies [14] and GPTZero, which assesses text for AI-generated patterns based on word choice and sentence length variability, though it falsely flagged 8 out of 10 human-written texts as AI-generated [15].

Text detection remains challenging due to the diverse, context-dependent nature of language, risking misinterpretation. To enhance reliability, analyzing the author's gender or personality based on the text and cross-referencing with existing information could be effective. Marking and detection technologies need to evolve for strengthened and trustworthy recorded media against all types of AI-generated content, authorized or not.

4.3. AI-Generated Content

The 'implied truth effect,' as discussed by Pennycook and Rand, highlights a counterintuitive outcome of selective labeling strategies: when not all false news stories are tagged, it inadvertently assigns higher credibility to those unlabeled, potentially increasing their dissemination [16]. This effect is significant in the realm of AI-generated content, which is addressed under Article 52(3) AIA. It mandates disclosure of AI-generated and deepfake content, except where such disclosure is limited by lawful or artistic reasons, and requires labeling of such texts in issues concerning public interest unless exempted for lawful purposes, human review, or editorial responsibility by natural persons.

However, these regulations face significant hurdles, such as low public awareness and skepticism towards labels, which could lead to misuse of exceptions for creative works and risk evasion of regulation through minimal compliance claims. The study's suggestion of 'Verified' tags for genuine content could combat these issues by focusing on affirming truths over tagging falsehoods, potentially counteracting the 'implied truth effect' in regulating deepfake and AI-generated content.

4.4. Digital Footprints and Online Spaces

The AIA's current definitions of 'biometric data' and 'publicly accessible spaces' (Articles 3(33), 3(39), and Recital 9b AIA) exclude digital footprints and online spaces, creating a regulatory gap that AI technologies could exploit to create deceptive online personas. This oversight is concerning given parental sharing of children's data can help algorithms build detailed digital profiles. The AIA does not adequately address the dynamic nature of digital identities or the emerging cybersecurity threats they pose.

The EU's commitment to data privacy is highlighted by the CJEU's Google Spain SL v. Agencia Española de Protección de Datos ruling, which established the 'right to be forgotten.' This ruling allows individuals to request the deletion of inaccurate, inadequate, irrelevant, or outdated personal information and classifies search engines' indexing and storage of personal data as 'data processing' subject to EU laws. This highlights the need for robust regulatory frameworks for managing personal data exposure within the digital ecosystem [17].

Additionally, the Data Protection Commissioner v. Facebook Ireland Limited and Maximillian Schrems case criticizes the Safe Harbor accord's inadequacy in protecting EU citizens from non-EU surveillance. This case highlights the critical need for stringent data protection standards to ensure data privacy across geopolitical boundaries, reflecting the EU's proactive stance in updating legislative frameworks to match technological advancements [18].

To address these issues, 'digital footprints' and 'online spaces' should be included within the AIA's regulatory scope. This revision would align with basic EU data protection principles, enhance protection against the complexities of global digital interconnectivity, and ensure the privacy and integrity of citizens in the digital age, keeping regulatory measures in step with technological advancements and the spread of digital identity.

4.5. Balancing Artistic Expression and Victim Rights

The phenomenon of deepfake content, such as Zubear Abdi's featuring Taylor Swift, presents a complex challenge. These deepfakes, protected under artistic expression and thus requiring minimal disclosure as per Article 52(3) AIA, often fail to mitigate the distress of individuals portrayed, forcing them to rely on broader privacy and defamation laws like GDPR and national legislation, which may not fully address their pain.

The GDPR sets out data processing principles and compensation rights for the damage from data breaches (Articles 4, 5, 6, 9, and 22 GDPR), hand in hand with the Digital Services Act (DSA), which calls for the responsibility and transparency of online platforms in relevant related sections (Articles 14, 15, 17, 19, 20, 26, 27, and 35 DSA). Moreover, the United Kingdom's Online Safety Act 2023 considers similar themes.

In response to high-profile incidents, the EU and the US are developing new legislation, such as the DEFIANCE Act, to tackle non-consensual synthetic media. However, these efforts struggle to balance anonymity and trust in digital spaces, as technological advances like blockchain may compromise the effectiveness of these laws.

To address these issues, we propose the following concrete amendments to the AIA:

- **Mandatory Disclosure Requirements:** All AI-generated content, including deepfakes, should have visible disclosure labels on social media, websites, and other digital media. Meta's recent initiative to label AI-generated content with "AI info" labels can serve as a model.
- **Specific Exemptions and Limitations:** Content that is defamatory, harmful, or misleading should have specific exemptions, requiring explicit consent for AI-generated likenesses of real individuals.
- **Victim Support and Redress Mechanisms:** Establish procedures for reporting and seeking redress for unauthorized use of likenesses in AI-generated content, including faster takedown processes, legal support, and compensation funded by AI content creators.
- **Enhanced Transparency Measures:** Platforms hosting AI-generated content should maintain transparency reports on the volume, types of AI-generated content, actions taken against harmful content, and compliance with disclosure requirements.
- **Technological Solutions for Verification:** Invest in technologies like digital watermarks and blockchain verification to authenticate and trace the origin of AI-generated content, distinguishing between legitimate artistic expressions and malicious deepfakes.

By implementing these concrete measures, the AIA can effectively balance the protection of artistic expression with the rights and well-being of individuals portrayed in AI-generated content. These amendments will enhance the regulatory framework's ability to address the complexities of modern digital media while safeguarding fundamental rights.

5. Virtual Influencers

Virtual influencers (VIs) are CGI-generated models designed to mimic human appearance and behavior, achieving high-level realism without physical existence in the real world [19]. An illustrative example is Miquela Sousa, better known as lilmiquela on Instagram (depicted in Figure 1), a 3D-rendered model that first debuted in 2016. What sets her apart is the underlying AI powering her persona, letting her make life-like animations and movements for social media engagements. But the AI makes the VIs much more accurate than the usual virtual characters do and even designs their manner of speaking and the nature of speech. This simulation includes both the manner of speaking and the nature of content shared with others through their respective social platforms. Therefore, this creates the persistence of a virtual identity equivalent to the consciousness but without true consciousness.

Figure 1: Instagram Profile of Miquela Sousa, a Virtual Influencer. Screenshot by Valerie Megan, via Instagram. (https://www.instagram.com/lilmiquela).

VIs are categorized mainly into anime-like (AVIs) and human-like virtual influencers (HVIs). Their development leverages AI to enhance realism and interaction with the real world. A study analyzed the believability of VIs through social media reactions, focusing on authenticity and user emotions. Utilizing methodologies like the Instagram Posts Extractor to gather engagement metrics and interactions from VI accounts, the Uncanny Valley Hypothesis has been confirmed [20]. This means VIs are almost human-like but not entirely, which can be unsettling to natural people.

VIs are equipped with advanced AI technologies, including Natural Language Processing, Machine Learning behavior prediction models, and CGI for visual realism, requiring large data inputs and sophisticated programming for seamless communication on social platforms. However, their human-like appearance and behavior pose unique legal challenges under the AIA, such as issues of consent, data privacy, and manipulation of human behavior.

Events like Miss AI, part of the World AI Creator Awards, highlight advancements in this field. While some implementations still appear unreal, studies into techniques like photoplethysmography, used in FakeCatcher, can possibly enhance their human-like quality. On the other hand, developing detection algorithms to identify such personas is crucial for protecting individuals from malicious AI-generated personas, focusing on improving element detection algorithms or adopting comprehensive detection methods.

6. Risk Assessment and Management

6.1. Criteria-Based Risk Classification

Article 6 (2a) AIA allows AI systems to avoid high-risk qualifications when they meet at least one of the four specified criteria. This approach raises concerns as it might overlook latent risks in unaddressed areas, therefore may render the risk assessment too lenient. To ensure comprehensive risk assessment and management, we propose the following concrete measures:

- **Multidimensional Risk Assessment Framework:** AI systems will be evaluated against all four criteria using advanced techniques like scenario analysis, fishbone method, causal mapping, Delphi technique, cross-impact analysis, bow tie analysis, and system-theoretic process analysis. These methods were exemplified in a recent study by Koessler and Schuett [21].

- **Quantitative Thresholds for Risk Evaluation:** Predefined thresholds for each criterion within the framework will be established, ensuring accurate reflection of AI system risks. Researchers and legislators will collaborate to set minimum thresholds for low-risk classification.
- **Robust Scoring Mechanism:** An overall risk score will aggregate the criteria scores. AI systems exceeding the cumulative score threshold will be classified as low-risk; those falling short will be deemed high-risk and undergo further investigation.

6.2. Learning from IMDA and NIST

The AIA is a significant step towards regulating AI development and deployment. However, the field of AI continues to evolve, and the AIA can benefit from learning from other countries' approaches to AI risk management.

Singapore's Infocomm Media Development Authority (IMDA) has introduced the Model AI Governance Framework, tailored to address Generative AI challenges. It emphasizes data integrity, security, content provenance, and incident reporting, incorporating nine dimensions of AI governance. This consultative approach involves stakeholders across policymakers, industries, and research communities, offering a model for the EU to ensure comprehensive governance that addresses both traditional AI risks and generative AI challenges [22].

The U.S. National Institute of Standards and Technology (NIST) developed the AI Risk Management Framework (AI RMF 1.0), focusing on continuous risk management throughout the AI system's lifecycle, organized under the core functions: Govern, Map, Measure, and Manage. This structured approach can help the EU establish a robust risk management culture within AI-developing organizations, enhancing the AIA's resilience and flexibility [23].

By integrating NIST's structured risk management and IMDA's generative AI focus, the EU can refine the AIA to foster innovation while mitigating potential risks associated with AI technologies.

7. Innovation and Regulatory Exemptions

According to Article 52c(-2) AIA, General Purpose AI (GPAI) models released under a free and open license are exempt from regulation unless associated with systemic risks. This exemption aims to foster collaboration and innovation, aligning with the AIA's objectives. However, the limitations on these exemptions may hinder innovation, especially for smaller enterprises [24, 25, 26]. Since scientific research enjoys comprehensive exemptions (Article 2(5a) AIA), it is reasonable to extend similar considerations to free and open-licensed AI systems, but only for the development phase, not deployment, to balance innovation with risk management.

8. Learning from the G7 Approach

The Hiroshima AI Process, established by the Group of 7 (G7), emphasizes safety, security, and democratic values in AI development, serving as an excellent model for enhancing the AIA [27]. Unlike the AIA's EU-centric risk management focus, the G7 framework promotes international collaboration and the convergence of global AI development on democratic values. Integrating G7's principle-based approach would foster a more innovation-friendly environment, ensuring AI's safety and trustworthiness while maintaining the EU's leadership in the global AI market.

The G7's strategy for interoperable AI systems can be adopted to encourage cross-border AI use, particularly in sectors like healthcare and transportation that require diverse legal governance. By incorporating these standards, the AIA would enhance data security and global data governance.

Incorporating G7 principles would elevate the AIA's standards, setting a global benchmark for AI regulation, and positioning the EU as a central figure in global AI governance. This integration would address challenges more effectively and promote technological fairness.

9. Conclusion

The AIA introduces a commendable risk-based framework for AI regulation, setting a precedent for future legislation within and beyond the EU. Its potential to align AI practices with societal norms and safety is vital for managing systemic risks.

However, the Act has substantial vulnerabilities that need addressing, especially in articles requiring high user awareness (e.g., Article 52 AIA). Not all users have the necessary technological literacy, weakening the Act's enforceability. The AIA's reliance on existing laws results in less stringent requirements for scenarios outside their scope, necessitating stricter standards to ensure accountability.

While this literature cannot exhaustively analyze the AIA, the highlighted issues indicate the need for a thorough legislative review. Legislators must scrutinize the Act for additional weaknesses to enhance its protective capabilities.

Our contributions analyze the AIA from both layman and researcher perspectives, providing solutions and insights, and examining global AI regulations. We also raise awareness about VIs and their potential threats. By refining the AIA, we aim to prevent misuse, promote innovation, and protect fundamental rights. Our targeted recommendations ensure the Act remains effective and responsive to future AI advancements.

10. Future Works

The AIA is a pioneering step in AI regulation, but ongoing research into its efficacy, implications, and areas for improvement is crucial. Researchers should analyze its real-world impacts, identify vulnerabilities, and propose innovative methods to enhance AI regulation. There's a significant need for developing robust AI marking and detection algorithms, considering ethical, social, and economic dimensions. The global significance of the AIA requires studies on international collaboration and harmonization of AI governance, engaging diverse stakeholders. The journey toward effective AI regulation is ongoing, and the AIA provides a foundational framework. Through research and collaboration, we can address challenges and harness AI's potential for positive societal contributions.

Acknowledgments

We would like to express our deepest gratitude to Dr. Bernhards "BB" Blumbergs for his expert insights and guidance in defense alliance.

References

[1] E. Parliament, Artificial intelligence act: Meps adopt landmark law, 2024. URL: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law.

[2] D. Hambling, Drones killing without oversight?, New Scientist 260 (2023) 8. doi:https://doi.org/10.1016/S0262-4079(23)01937-1.

[3] J. Wagstaff, New model army, 2023. URL: https://www.imf.org/en/Publications/fandd/issues/2023/12/Case-Studies-New-model-army-Jeremy-Wagstaff.

[4] D. A. R. P. A. (DARPA), Darpa initiates design of longshot unmanned air vehicle, 2021. URL: https://www.darpa.mil/news-events/2021-02-08.

[5] G. Rak, 11 intriguing engineering milestones to look for in 2024, 2024. URL: https://spectrum.ieee.org/technology-in-2024.

[6] NATO, Summary of NATO's Autonomy Implementation Plan, 2022. URL: https://www.nato.int/cps/en/natohq/official_texts_208376.htm.

[7] T. Nakamura, European constitutional treaty: Commentary and translation, House of Representatives Constitutional Investigation Committee Office, 2004. URL: https://www.shugiin.go.jp/internet/itdb_kenpou.nsf/html/kenpou/chosa/shukenshi056.pdf/$File/shukenshi056.pdf.
[8] S. Lloyd, C. Weedbrook, Quantum generative adversarial learning, Phys. Rev. Lett. 121 (2018) 040502. doi:10.1103/PhysRevLett.121.040502.
[9] U. A. Ciftci, I. Demir, L. Yin, Fakecatcher: Detection of synthetic portrait videos using biological signals, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020). doi:10.1109/TPAMI.2020.3009287.
[10] E. Prashnani, M. Goebel, B. S. Manjunath, Generalizable deepfake detection with phase-based motion analysis, 2022. arXiv:2211.09363.
[11] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization, 2023. arXiv:2212.10957.
[12] P. G. Traynor, K. Butler, L. E. Blue, L. Vargas, K. S. Warren, H. Abdullah, C. Gibson, J. N. Odell, Detecting deep-fake audio through vocal tract reconstruction, 2020. URL: https://patents.google.com/patent/US20220036904A1/en.
[13] A. Pianese, D. Cozzolino, G. Poggi, L. Verdoliva, Deepfake audio detection by speaker verification, 2022. arXiv:2209.14098.
[14] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models, 2023. arXiv:2301.10226.
[15] M. Hutson, Detection stays one step ahead of Deepfakes—for now, 2023. URL: https://spectrum.ieee.org/deepfake.
[16] G. Pennycook, D. G. Rand, The psychology of fake news, Trends in Cognitive Sciences 25 (2021) 388–402. doi:https://doi.org/10.1016/j.tics.2021.02.007.
[17] C. of Justice of the European Union, Google spain sl, google inc. v. agencia española de protección de datos (aepd), mario costeja gonzález, Case C-131/12, ECLI:EU:C:2014:317, 2014.
[18] C. of Justice of the European Union, Data protection commissioner v facebook ireland limited and maximillian schrems, Court of Justice of the European Union, C-311/18, 2020.
[19] M. Conti, J. Gathani, P. P. Tricomi, Virtual influencers in online social media, IEEE Communications Magazine 60 (2022) 86–91. doi:10.1109/MCOM.001.2100786.
[20] J. Arsenyan, A. Mirowska, Almost human? a comparative case study on the social media presence of virtual influencers, International Journal of Human-Computer Studies 155 (2021) 102694. doi:https://doi.org/10.1016/j.ijhcs.2021.102694.
[21] L. Koessler, J. Schuett, Risk assessment at agi companies: A review of popular risk assessment techniques from other safety-critical industries, 2023. URL: https://arxiv.org/abs/2307.08823.
[22] I. M. D. Authority, Singapore proposes framework to foster trusted Generative AI development, 2024. URL: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/public-consult-model-ai-governance-framework-genai.
[23] N. I. of Standards, Technology, AI Risk Management Framework, 2024. URL: https://www.nist.gov/itl/ai-risk-management-framework.
[24] P. Cihon, How to get ai regulation right for open source, 2024. URL: https://github.blog/2023-07-26-how-to-get-ai-regulation-right-for-open-source/.
[25] A. Prabhakar, The eu ai act is a cautionary tale in open-source ai regulation, 2023. URL: https://datainnovation.org/2023/11/the-eu-ai-act-is-a-cautionary-tale-in-open-source-ai-regulation/.
[26] P. Keller, How will the ai act deal with open source ai systems?, 2022. URL: https://openfuture.eu/blog/how-will-the-ai-act-deal-with-open-source-ai-systems/.
[27] Group of Seven, G7 leaders' statement, 2023. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2023/12/06/g7-leaders-statement/.