The legal implications of AI transparency;
When content and data traceability measures meet human rights law

Keywords: Transparency, traceability, synthetic content, international law, watermarking

Names: Letrone William, (ルトロン　ウイリアム), Hayashi Mika (林　美香)

1 Objective

This research aims to raise awareness for the legal implications of traceability measures for AI-generated content. Indeed, a plethora of actors are currently contemplating the implementation of traceability methods, including watermarking, with a view to mitigating the risks emanating from generative AI. Although greater traceability levels could really benefit intellectual property rights and information integrity, it also presents certain risks. In fact, some initiatives introducing content traceability methods as a fix for AI-enabled issues have been met with skepticism. This was especially the case for digital watermarking, which remains a contentious area of research on its own. Critics warned about potential unwanted side-effects on datasets diversity, knowledge dissemination, and most importantly, the human rights of free speech and privacy. Hence, while traceability mechanisms for generative AI should be praised for their overall contribution to transparency, which has been elevated by the European High-Level Expert Group as a mandatory prerequisite to achieve AI trustworthy AI, outmost care should be exhibited by regulators tackling traceability in the context of AI-generated content and AI training data. This research attempts to explore the dichotomy surrounding AI transparency, showing that while increased transparency through traceability mechanisms can enhance accountability and trust, it may also inadvertently create tensions with human rights.

2 Methods

The research combines elements of evaluative and comparative legal research methods. It is evaluative in the sense that it mobilizes key legal rules to explore how they interface with traceability tools. It also features aspects of the comparative legal methodology in that it provides a comparison of existing laws establishing content labelling requirements. The arguments and ideas contained in the research draw from the relevant case law and literature from the fields of law and computer science.

3 Results

The research covered, in turn, AI transparency as mandated by the Law, the positive contribution of content and data traceability measures in terms of privacy and transparency, and the human rights risks associated with some of them. In doing so, the research offered helpful legal guidance to computer scientists seeking to develop content and data traceability tools at scale.

4 Conclusion

It is now well-documented that generative AI has the potential to exacerbate information disorders. There is also a growing awareness that generative AI might contribute to copyright violations. As the world is entering a period of intense regulatory activity around artificial intelligence, it is crucial that the laws are designed so that the pursuit of transparency in AI does not clash with human rights. By exploring the concepts of AI transparency and traceability, and by applying a legal perspective on non-invasive methods such as metadata-based labelling, and more sophisticated ones such as digital watermarking, the research has demonstrated both the legal utility and limits of content and data traceability.

Main references

- Center for Democracy and Technology, (CDT), “Privacy Principles for Digital Watermarking” (2008).
- Fernandez, P., Level A., & Furon T., “What Lies Ahead for Generative AI Watermarking”, *2nd Workshop on Generative AI and Law,* co-located with the International Conference on Machine Learning Vienna, Austria, (2024).
- Henderson P., “Should the United States or the European Union Follow China’s Lead and Require Watermarks for Generative AI?”, *Georgetown Journal of International Affairs,* (2023).
- Jaidka K. et al., “Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy”, *Digital Government: Research and Practice*, 6(1), Article 11, (2025).
- Jakesch M., Hancock J.T., & Naaman M., “Human heuristics for AI-generated language are flawed”, *PNAS*, 120(11), (2023).
- Knott, A., Pedreschi, D., Jitsuzumi, T. et al., “AI content detection in the emerging information ecosystem: new obligations for media and tech companies”, *Ethics and Information Technology*, 26(63), (2024).
- Knott, A., Pedreschi, D., Chatila, R. et al*.,* “Generative AI models should include detection mechanisms as a condition for public release”, *Ethics and Information Technology* 25(55), (2023).
- Wang T., Zhang Y., Qi S., Zhao R., Xia Z., & Weng J., “Security and Privacy on Generative Data in AIGC: A Survey” *ACM Computer Survey* 57(4), Article 82, (2025).