Human oversight in Automated Decision-Making: A Legal Safeguard Against Algorithmic Discrimination Under EU Law (Working Title)

Davide Baldini [a], [1]

[a]Università degli Studi di Firenze | Maastricht University (double PhD degree)
ORCID: https://orcid.org/0009-0009-9854-1019

Abstract. This thesis critically examines the role of human oversight as a legal safeguard against algorithmic discrimination in automated decision-making systems under European Union law. With AI increasingly employed in crucial domains such as employment, finance, and public administration, algorithmic biases pose significant risks to fundamental rights, especially non-discrimination. While EU regulations like the GDPR, AI Act, and Platform Workers Directive mandate human oversight to mitigate these risks, empirical evidence highlights substantial cognitive and practical challenges limiting oversight's effectiveness, such as automation bias and difficulties interpreting algorithmic decisions. This interdisciplinary research integrates doctrinal legal analysis, empirical case studies, cognitive psychology, and insights from human-computer interaction literature to propose an original framework for genuinely meaningful oversight. The findings underscore the necessity for targeted strategies – including specialized bias training, explicit intervention guidelines, and enhanced AI explainability protocols – combined with innovative regulatory instruments like AI regulatory sandboxes[2] and technical standards. By systematically addressing oversight's documented limitations and providing practical recommendations for policymakers, regulators, and AI developers, this research contributes significantly to ensuring responsible, effective, and legally compliant human oversight in automated decision-making across the EU digital landscape. This research also builds on the author's previously published work, including analyses of Article 22 GDPR and the right to explanation[3], as well as studies on the regulatory potential of AI sandboxes under EU digital law[4]. These publications form the theoretical foundation for the dissertation, while the present project extends them by integrating empirical findings and interdisciplinary insights into a comprehensive framework for meaningful human oversight.

1. Research Questions

This dissertation seeks to systematically investigate the main research question and sub-questions indicated below:

Main Research Question:

- *How can human oversight mechanisms mandated by EU secondary law be structured and implemented effectively to mitigate algorithmic discrimination?*

Sub-questions:

- *What are the specific human oversight requirements established under EU secondary law (GDPR/LED, AI Act, Platform Workers Directive), and how do they interact and overlap?*
- *What practical limitations affect the effectiveness of human oversight, particularly concerning cognitive biases and epistemological differences between human and algorithmic reasoning?*

[1] Corresponding Author. Email: davide.baldini@unifi.it | davide.baldini@maastrichtuniversity.nl.

[2] Regulatory sandboxes may be generally defined as "controlled environments in which novel technological applications can be tested under the close supervision of competent authorities for a limited time and under defined conditions". In this project, the author specifically refers to "AI Regulatory Sandboxes", namely the regulatory sandboxes as mandated under Chapter V of the AI Act.

[3] Baldini, D., 'Article 22 GDPR and prohibition of discrimination. An outdated provision?', CyberLaws, 20 August 2019 <https://www.cyberlaws.it/2019/article-22-gdpr-and-prohibition-of-discrimination-an-outdated-provision/> [accessed 11 September 2025]; Dirutigliano, J., and Baldini, D., 'The Right to Explanation: Legal Challenges and the Future of Fairness in Automated Decision-Making', Journal of AI Law and Regulation, 2.2 (2025), pp. 120–38, doi:10.21552/aire/2025/2/5..

[4] Baldini, D., and Francis, K., 'AI Regulatory Sandboxes between the AI Act and the GDPR: The Role of Data Protection as a Corporate Social Responsibility' (CEUR Workshop Proceedings, 2024) <https://flore.unifi.it/handle/2158/1372612> [accessed 28 March 2025]; • Bagni, F., and others, 'White Paper on Regulatory Sandboxes for AI and Cybersecurity', SSRN Scholarly Paper no. 5268812 (Social Science Research Network, 4 February 2025) <https://papers.ssrn.com/abstract=5268812> [accessed 11 September 2025].

- *What structural and procedural measures (e.g., training, guidelines, explainability protocols) enhance human oversight's effectiveness in mitigating discrimination?*
- *How can regulatory tools such as AI regulatory sandboxes, technical standards, and common specifications under the AI Act facilitate effective oversight implementation without compromising AI innovation and efficiency?*

Through answering these questions, this dissertation aims to offer a nuanced and empirically grounded framework for the legally mandated human oversight mechanisms, advancing both theoretical understanding and practical implementation within EU digital law. This contribution addresses a significant gap by synthesizing legal scholarship with cognitive and technical insights, thus ensuring that human oversight effectively fulfils its intended role as a safeguard against algorithmic discrimination.

2. Preliminary findings

Early findings from the research, based on insights from existing literature and research, affirm both the importance of human oversight in theory and the complexity of making it effective in practice. The legal analysis reveals that EU laws increasingly stress "meaningful" human oversight as a safeguard, but they vary in specificity. For example, Article 22 GDPR affords individuals a right to human intervention in fully automated decisions, yet offers scant detail on how that oversight should be executed. In contrast, the proposed AI Act (Article 14) and the Platform Workers Directive (Article 10) explicitly require human oversight for certain high-risk AI systems and provide more concrete implementation criteria. This evolution signals the EU's growing resolve to rely on human oversight as a regulatory tool, while also hinting at potential overlaps and inconsistencies between instruments that the thesis is mapping out. At the same time, the interdisciplinary inquiry underscores significant limitations of human oversight as currently practiced. Empirical case reviews and literature indicate that human overseers often struggle to detect subtle or systematic biases in AI outputs. Cognitive challenges – like human's difficulty in interpreting AI's probabilistic reasoning or maintaining vigilance during high-volume automated processes – mean that oversight can devolve into a perfunctory "rubber-stamping" exercise. Indeed, preliminary analysis suggests that untrained or unsupported human oversight may sometimes give a false sense of security (what Brennan-Marquez et al., 2019 have called an "apparent" rather than actual human involvement in ADM, namely where oversight is merely formal and does not affect algorithmic outcomes) without actually reducing discriminatory outcomes.

These insights reinforce the thesis' central problem: without deliberate design and support, human oversight alone will not reliably prevent algorithmic biases. The research has also identified promising strategies to enhance oversight. A review of emerging recommendations shows consensus around measures like bias awareness training for human reviewers, well-defined intervention protocols (when and how overseers should intervene on an AI's decision), and improved explainability of AI decisions to aid human understanding. Such measures are hypothesized to mitigate issues like automation bias and information asymmetry, making oversight more effective.

Additionally, initial exploration of regulatory innovations indicates that tools like AI regulatory sandboxes and technical standards could play a role in refining oversight practices. AI regulatory sandboxes (as envisioned in the AI Act) would allow controlled experimentation with oversight mechanisms in diverse contexts, helping stakeholders learn what works before wider deployment. Common standards or guidelines could disseminate best practices and ensure a baseline quality of oversight across industry. However, the preliminary analysis also cautions that these instruments are not panaceas: sandbox approaches must avoid regulatory capture or fragmentation, and overly rigid standards might stifle context-specific adaptations. These nuanced findings will inform the dissertation's proposed framework, stressing that effective oversight requires both multi-faceted human training and support, and adaptive governance mechanisms that encourage innovation while upholding rights.

3. Directions for the remaining work

Going forward, the research will delve deeper into evaluating and validating the proposed oversight framework. This includes refining the normative criteria for "meaningful" oversight and testing the feasibility of recommended interventions (for instance, via hypothetical scenarios or further case studies). The upcoming thesis chapters will formulate concrete policy recommendations for EU institutions and practical guidelines for organizations implementing ADM oversight.

The dissertation's main contribution will be the development of a comprehensive oversight framework. Its preliminary structure consists of:

- Setting out and systematizing the (currently scattered) normative criteria aimed at ensuring genuinely "meaningful" human intervention is achieved, as derived from a critical examination of relevant EU law instruments including GDPR, LED, AI Act, DSA, and Platform Workers Directive, as interpreted by relevant case-law and soft law instruments such as the EDPB guidelines and the HLEG Ethics Guidelines for Trustworthy AI. This will enable to conceptualize overlapping human oversight requirements in EU law, to be then leveraged as a compass through the rest if the dissertation.

- Accounting for the practical challenges that stand in the way of the legally-mandated standard of "meaningful" human oversight, drawing upon relevant literature (e.g., on cognitive psychology and human-machine interaction) and empirical research, in order to highlight structural and cognitive limitations. This part will focus particularly on cognitive biases affecting human oversight, such as automation bias and algorithm aversion. Additionally, this chapter explores fundamental epistemological differences between qualitative human reasoning and quantitative algorithmic processing. Empirical insights and scholarly contributions from relevant literature are discussed to illustrate these issues clearly, acknowledging complementary insights from fields such as cognitive psychology and behavioural economics.

- Establishing potential strategies to enhance human oversight, focusing on innovative regulatory and institutional approaches and building on the interaction between the two previous points. This part will present normative criteria essential to establish genuinely meaningful oversight, including accountability, transparency, autonomy, and fairness. This part will also investigates AI regulatory sandboxes as a possible instrument to address the possible pitfalls and shortcomings of human oversight, as singled-out in the preceding sections, examining their role in practically supporting context-sensitive, iterative implementations of human oversight in high-risk use-cases. The aim of this section is to lay down practical recommendations for regulators, providing suggestions to strengthen the overall effectiveness of oversight mechanisms foreseen by the growing corpus of EU digital law.

In this respect, Participation in the Doctoral Consortium will be valuable at this stage. Presenting these preliminary insights to an interdisciplinary audience of AI researchers, ethicists, and legal experts, may allow the gathering of constructive feedback to strengthen the analysis and ensure the work's relevance beyond the legal domain. The Consortium's input can help refine the framework's assumptions, suggest new angles (e.g., HCI design improvements or comparative perspectives).

4. Methodology

This research employs, primarily, a legal-doctrinal approach supplemented by a legal-theoretical approach (also drawing from other relevant fields, especially cognitive psychology) as to its contextualization and reflection, with a view to address the normative, legal, and practical dimensions associated with the concept of human oversight in AI systems. Specifically, this research relies on a combination of doctrinal legal analysis, case study and soft law analysis, as well as interdisciplinary analysis. This multi-dimensional methodological approach is necessary to ensure comprehensive coverage of the research questions and the robust exploration of possible oversight mechanisms.

4. References

- European Agency for Fundamental Rights 'Bias in Algorithms - Artificial Intelligence and Discrimination | European Union Agency for Fundamental Rights', 29 November 2022 https://fra.europa.eu/en/publication/2022/bias-algorithm
- Baldini, D., 'Article 22 GDPR and prohibition of discrimination. An outdated provision?', CyberLaws, 20 August 2019 <https://www.cyberlaws.it/2019/article-22-gdpr-and-prohibition-of-discrimination-an-outdated-provision/> [accessed 28 March 2025].
- Baldini, D., and Francis, K., 'AI Regulatory Sandboxes between the AI Act and the GDPR: The Role of Data Protection as a Corporate Social Responsibility' (CEUR Workshop Proceedings, 2024) <https://flore.unifi.it/handle/2158/1372612> [accessed 28 March 2025];
- Bagni, F., and others, 'White Paper on Regulatory Sandboxes for AI and Cybersecurity', SSRN Scholarly Paper no. 5268812 (Social Science Research Network, 4 February 2025) <https://papers.ssrn.com/abstract=5268812> [accessed 11 September 2025].
- Brennan-Marquez, Kiel, Daniel Susser, and Karen Levy, 'Strange Loops: Apparent versus Actual Human Involvement in Automated Decision-Making' (Social Science Research Network, 2 October 2019) <https://papers.ssrn.com/abstract=3462901>
- Constantino, Jorge, 'Exploring Article 14 of the EU AI Proposal: Accountability Challenges of the Human in the Loop When Supervising High-Risk AI Systems in Public Administration' (Social Science Research Network, 17 August 2022) https://papers.ssrn.com/abstract=4254940
- Dirutigliano, J., and Baldini, D., 'The Right to Explanation: Legal Challenges and the Future of Fairness in Automated Decision-Making', Journal of AI Law and Regulation, 2.2 (2025), pp. 120–38, doi:10.21552/aire/2025/2/5.
- 'Explanatory Memorandum on the Updated OECD Definition of an AI System', OECD, 4 March 2024 <https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html>
- Fink, Melanie, 'Human Oversight under Article 14 of the EU AI Act' (Social Science Research Network, 14 February 2025) <https://papers.ssrn.com/abstract=5147196>
- Green, Ben, 'The Flaws of Policies Requiring Human Oversight of Government Algorithms', Computer Law & Security Review, 45 (2022), p. 105681, doi:10.1016/j.clsr.2022.105681
- Panezi, Argyri, 'Requirements of High-Risk AI Systems: AI Act. Article 14. Human Oversight' (Social Science Research Network, 1 July 2024) <https://papers.ssrn.com/abstract=5131229>
- Panigutti, Cecilia, and others, 'The Role of Explainable AI in the Context of the AI Act', in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23 (Association for Computing Machinery, 2023), pp. 1139–50, doi:10.1145/3593013.3594069
- Solove, Daniel J., and Hideyuki Matsumi, 'AI, Algorithms, and Awful Humans' (Social Science Research Network, 16 October 2023) <https://papers.ssrn.com/abstract=4603992>
- Gaudeo, A., et al., 'Understanding the Impact of Human-AI Interaction on Discrimination - European Commission', 25 February 2025 <https://policy-lab.ec.europa.eu/news/understanding-impact-human-ai-interaction-discrimination-2025-01-10_en>
- Wachter, Sandra, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law' (Social Science Research Network, 15 February 2022), doi:10.2139/ssrn.4099100
- Xenidis, Raphaële, and Linda Senden, 'EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination' (Social Science Research Network, 30 September 2019) <https://papers.ssrn.com/abstract=3529524>