The Human Oversight Fallacy–and What We Can Do About It

Angelica Henestrosa[1,*], Alexeï Grinbaum[2] and Ivan Yamshchikov[1]

[1]CAIRO, Technische Hochschule Würzburg-Schweinfurt (THWS), Franz-Horn-Straße 2, 97082 Würzburg, Germany
[2]Commissariat a L'Energie Atomique et aux Energies Alternatives (CEA-Saclay/LARSIM), 91191 Gif-sur-Yvette Cedex, France

Abstract

Human oversight, as one of several ethical principles required for the safe use of AI systems, plays a prominent role in the European Union's Artificial Intelligence Act (AIA). This position paper argues that, while human oversight remains a necessary safeguard to mitigate the risks posed by AI systems aimed at directly interacting with humans across diverse domains, its current conceptualization rests on problematic assumptions about human judgment, epistemic access, and institutional responsibilities. Drawing on empirical findings from psychology and human–AI interaction research, we show that human oversight, as presently envisaged, runs risk to remaining an illusion of control if the conditions for meaningful oversight are not made explicit. To address these limitations, we propose a conceptual reorientation of human oversight away from reactive approval and toward structurally embedded control, by suggesting incentive-based ethical stress testing, the establishment of normative clarity, and the systematic accounting for the limits of human oversight. We conclude that only by grounding oversight in realistic assumptions about human cognition and institutional incentives can it fulfill its intended role in ensuring trustworthy AI.

Keywords

Human Oversight, Human Supervision, Ethical AI, European AI Act

1. Introduction

Human oversight is commonly presented as a primary mechanism through which humans retain control over increasingly capable artificial intelligence (AI) systems: humans define objectives for AI systems, choose and deploy technical solutions to reach these objectives, and monitor their operation. Human oversight is the first of the seven key requirements for trustworthy AI promoted by the European Commission [7]. In this context, human oversight is understood as an active involvement of at least one human operator who monitors automated decision-making systems, evaluates their outputs, and retains the ability to intervene where necessary [28]. The EU Artificial Intelligence Act (AIA) further operationalizes this principle, requiring that "proper oversight mechanisms" be ensured through human-in-the-loop, human-on-the-loop, or human-in-command approaches, with the explicit aim of preventing or minimizing risks to health, safety, and fundamental rights (AIA article 14 and recital 27). To this end, the AIA requires that a responsible human person remains in control and maintains authority over the AI system (AIA recital 48).

While this regulatory framework articulates a normatively appealing vision of human control, it rests on strong assumptions regarding human cognitive capacities, the epistemic access to complex AI systems and effective control over them. As the AIA regulations are approaching implementation, the question arises as to their meaningful realization in the actual human-AI interaction. In particular, the validity and effectiveness of the envisaged human oversight need to be empirically proved.

These concerns become particularly relevant in the context of Large Language Models (LLMs) and other forms of general purpose AI (GPAI) designed for direct interaction with end users, without a stable intermediary and across diverse application areas. The question arises as to whether human oversight,

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*Corresponding author.
angelica.henestrosa@thws.de (A. Henestrosa); Alexei.Grinbaum@cea.fr (A. Grinbaum); ivan.yamshchikov@thws.de (I. Yamshchikov)
0000-0002-0733-9045 (A. Henestrosa); 0000-0002-7484-1553 (A. Grinbaum); 0000-0003-3784-0671 (I. Yamshchikov)

as currently envisaged in the AIA, may be already insufficient given the current state of technology. While cognitive risks for humans emerge gradually, ethical, psychological, and societal consequences and effects are not foreseeable. Consequently, they cannot be controlled via human oversight here and now.

Against this background, this position paper identifies three structural weaknesses in the prevailing approaches to human oversight and argues for a conceptual reorientation of this ethical principle, by outlining several constructive directions of risk mitigation.

2. Thesis Statement

We submit that current approaches in AI governance fall short in meeting several fundamental objectives of human oversight of AI systems. These arise due to: (i) the systemic fallibility of human judgment, (ii) practical limitations of how human oversight is implemented and currently envisioned, and (iii) conceptual ambiguities of meaningful oversight. Essentially, human oversight achieves only one key objective of assigning responsibility to persons rather technical systems.

Against this background of limitations, the paper argues for redesigning and reducing oversight on the basis of empirically grounded limits of human judgment. We propose to move human oversight away from being an ultimate judge to being an ultimate bearer of responsibility, while effective control and actual decision-making should be embedded in a broader, multi-layered operational framework.

3. Human Oversight as an Illusion of Control

In the following, three illustrative aspects are presented that demonstrate why human oversight, as it is currently envisaged, risks functioning as an illusion of control over AI systems. Using these examples, we show how assumptions about human judgment, organizational responsibility, and epistemic access undermine the effectiveness of oversight in practice and may fail precisely where it is intended to safeguard.

3.1. Human Oversight as Anthropocentric Overconfidence

The widespread reliance on human oversight as a cornerstone safeguard against AI risks reflects a form of anthropocentric thinking: the assumption that any control over human action should belong to human actors. This smells of overconfidence: merely inserting humans into complex decision-making loops may not be sufficient to mitigate systemic, epistemic, and ethical failures. This is because human judgment is systematically biased, context-dependent, and readily shaped by AI influences and cues [30, 22, 10]. Human cognition is not a separate cognitive system: it increasingly forms a subsystem of a human-AI interacting whole.

Prevailing definitions of human oversight implicitly rely on a self-image of humans as critical and discerning evaluators, yet they insufficiently account for well-documented cognitive and structural biases that render oversight error-prone. This includes the tendency to rely on information that aligns with one's own values and believes [26, 23], but also phenomena reflecting a reduced level of critical engagement in the presence of highly reliable systems and paying insufficient attention to contradictory evidence, known as automation bias [24, 27], or the control problem [33].

Empirical evidence suggests that this self-image does not survive actual confrontation with AI outputs. People have prior beliefs and assumptions about who performs a task better [20], often over-relying on humans, even though the AI would outperform human performance in a task [2, 16]. However, when interacting with AI systems, people frequently perceive them as equally credible [21, 32]. Moreover, users are typically unable to correctly decide when to diverge from algorithmic recommendations [12, 7]. Crucially, these limitations are not confined to inexperienced users but are also observed among domain experts [5]. Thus, precisely in situations when human oversight is supposed to function, it might be severely limited.

3.2. Human Oversight as Ethical Silver Bullet

Currently, there is no shared operational definition of what constitutes meaningful human oversight besides the consensus that it should not be rubber stamping [11]. At the same time, it remains unclear where exactly oversight is legally required, nor does the current version of the EU AIA provide specific guidelines or standards for its implementation. For example, start-ups and AI developers typically possess substantial technical expertise, yet they lack meaningful responsibility for the long-term societal consequences of the use of AI systems. Under these circumstances, oversight appears to be promoted as a procedural and regulatory compensation for unresolved legal tensions around responsibility.

At an organizational level, human supervisors might be used as foci of liability in the absence of commonly accepted regulatory mechanisms governing AI accountability [31]. On this view, oversight is necessary, not because it is effective but because it reassigns responsibility to persons rather than technical systems. This lack of effectiveness is also linked to the question of when exactly human oversight is considered to be fulfilled and how this can be assessed [18, 6]. As a result, current regulatory approaches risk removing responsibility from the production chain of AI outputs and placing it on the supervisor rather than the provider, the operator, or the user. At the end of the chain, the actual users are the ones bearing the consequences of oversight, but a claim of inadequate oversight may absolve them of their own responsibility. This seems to be a quick shortcut for the users who lack the necessary epistemic conditions and institutional support to distribute responsibility in a better grounded way.

At the same time, human oversight that lacks the effective ability to contest AI-influenced user decisions functions as a protector of the status quo. This may put users at risk even in the presence of oversight. Recent examples of the dissemination of potentially illegal content [4] and suicidal behavior following human-AI interactions [13, 29] demonstrate how simply warning against misuse or excluding it in the terms and conditions is not enough.

3.3. Human Oversight is Symbolic and Epistemically Weak

At the moment, human oversight often functions primarily as a means of satisfying regulatory requirements rather than as a mechanism that substantively improves the quality of human decisions based on AI outputs. Expert access to the technology is not guaranteed for the supervisors, oftentimes due to proprietary architectures, closed-source models, or insufficient transparency about the AI system's behavior. A meaningful question to ask from both the regulatory and the technical perspectives is whether human supervisors are capable of reasonably contradicting AI outputs despite cognitive influences on their own thinking, and whether they can demonstrably improve the quality of human-AI interaction. With AI systems growing increasingly complex and opaque, supervisors need ever more expertise and control [28, 14].

Recently, Sam Altman argued that restrictions on ChatGPT's functionality in order to protect users with mental health-related issues would make it less usable and enjoyable for many other users to interact with the chatbot[1], yet stating they had mitigated serious mental health related issues. At the same time, especially in the context of GPAI, it is oftentimes called for AI literacy to enable users to use these tools responsibly. However, GPAI are not merely open-ended tools, whose safe use can be ensured through user education, adequate instructions, and the mitigation of the most obvious failure modes. They are products deployed at scale into broad societal contexts which is why providers are obliged to prevent or mitigate abusive use ex ante. Within the current regulatory landscape, this obligation remains insufficiently specified, particularly for AI systems that fall outside the high-risk classification of the EU AIA. This creates ambiguity regarding the boundaries of permissible functionality, weakening both accountability and governance.

[1] https://x.com/sama/status/1978129344598827128; Accessed on January 16, 2026.

4. Reclaiming Meaningful Human Control

We argue that current efforts to achieve successful human oversight, especially in GPAI systems, are insufficient to fulfill their ethical and societal responsibility. Importantly, however, it should be emphasized that the weaknesses identified can be mitigated at least partially.

In the following, we outline three recommendations to address these limitations and to equip human oversight with the technical and conceptual instruments required to function as a meaningful mechanism of control and impact on AI systems.

4.1. Incentive-Based Ethical Stress-Testing

The current incident with Grok illustrates some of the structural weaknesses of current oversight approaches and, above all, the obstacles in the prevention of AI abuse [25]. Here, exemplarily, the legal constraints prohibit developers and red teams from systematically testing AI systems for harmful behavior with regard to child abuse material, classifying this already as a criminal offense. As a result, measures to counteract abusive use cannot be tested at an early stage.

More generally, red teaming is a mechanism to target not only technical model vulnerabilities but also predictable failures of human oversight, which go beyond hacking attacks and jailbreak attempts [9, 8]. By engaging multidisciplinary red teams consisting of domain experts, developers, psychologists, and ethicists, for instance, various perspectives can be merged considering immediate risks and potential long-term consequences of failed oversight.

However, effective red teaming cannot be achieved by solely imposing formal obligations, as such requirements do not guarantee the quality and independence of testing. Instead, testing the riskiest and ethically questionable applications should be externalized and embedded in an incentive-compatible framework. One approach could be to encourage independent researchers, civil society organizations, and users to conduct and report targeted tests of AI systems in order to formally report them. Regulators could support this by establishing protected report channels, standardized disclosure formats, and the legal requirements for good-faith testing. Furthermore, financial rewards or tax benefits could be introduced to incentivize the developers as well. Such an approach would align safety objectives with market incentives and embed them legally, both the most important levers in AI regulation.

4.2. Establishing Normative Clarity

Referring back to previous examples, clarity regarding unauthorized use is lacking, which concerns oversight as well. In the absence of sufficiently specific regulations, competitors currently orient themselves towards major market players for guidance when it comes to evaluating their own compliance decisions, potentially creating a race to the bottom dynamic. Therefore, the agreement on minimum standards across competitors is essential. This does not work without regulatory clarity on the boundaries of AI systems that must be legally and technically excluded in advance. Here, properly designed oversight can function as an anchor mechanism, protecting these non-negotiable boundaries, and preventing competitive pressure from eroding safety standards.

To address this gap, explicit normative boundaries should be established, which cannot be overridden by human oversight. Such could take the form of hard normative stops, preventing certain model behaviors regardless of contextual judgments about utility or benefit. In addition, providers of GPAI systems could be required to publish mandatory normative use declarations which explicitly define acceptable, restricted, and prohibited uses. Consequently, normative clarity would reduce reliance on case-by-case human oversight that is not suitable for monitoring major societal impacts.

4.3. Accounting for the Limits of Human Judgment

Without explicit recognition of the contexts in which human oversight is likely to fail, oversight itself can become a vulnerability and an entry point for security-compromising actions [3]. Thus, this paper argues for a capability-aware oversight approach, which actively de-biases human judgment and

supervision by identifying its inherent weaknesses and limitations. To this end, AI providers should adopt proportionate and evidence-based debiasing interventions aligned with the most recent and state-of-the-art research on automation bias [19] or on the risk of over-reliance induced by certain forms of explanations [15, 17], for instance. However, current regulations aiming for transparency and explainability are rather optimized for compliance, not for human cognition. This requires a combination of several diverse measures, each of which addresses different weaknesses and limitations of oversight.

Accounting for human cognitive biases requires, first, the systematic identification of conditions under which such biases are most likely to arise, such as repetition fatigue or prolonged exposure to highly reliable systems. On this basis, the circumstances under which human supervisors are expected to contradict system outputs should be made explicit and operational. Mechanisms commonly invoked to support oversight, including transparency and explainability, must themselves be cognitively aligned, acknowledging that in certain contexts they can increase susceptibility to error or over-reliance rather than mitigate it [1].

Once the limits of human oversight have been established, defining non-oversight zones is possible so that oversight can be targeted only where humans can add epistemic value, ideally in a collective rather than individual manner. Finally, effective oversight necessitates dedicated meta-oversight structures that regularly evaluate these rules and thus can trigger redesign where oversight effectiveness demonstrably degrades.

Taken together, this would entail a broader shift in perspective: oversight is most effective when its limits are clearly outlined and when it is supported by non-human safeguards rather than used to compensate for their absence, thus being one of several components in a multi-layered structure.

5. Conclusion

This position paper does not seek to abolish human oversight as an ethical principle and necessary governance instrument. On the contrary, it argues for its preservation through a substantial reconfiguration of its currently envisaged form. Rather than treating human oversight as a reactive safeguard against emerging technical risks, the conditions for successful oversight must be clarified in advance. We argue that misinterpreting human oversight as a remedy for currently unsolved problems risks both ethical complacency and practical failure. A more promising approach lies in designing oversight regimes that are grounded in what we know about human judgment and cognition and that combine this with a regulatory framework that creates the prerequisites for implementing and evaluating successful monitoring. Underpinned with appropriate measures and a clearly defined scope can human oversight meaningfully contribute to the ethical, safe, and responsible deployment of advanced AI systems.

Acknowledgments

This research is supported by AIOLIA project funded by the European Commission under Grant Agreement 101187937.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL and GPT-5.2 in order to: Grammar, spelling check, rephrasing for readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 78–91. ACM. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534164. URL https://dl.acm.org/doi/10.1145/3514094.3534164.

[2] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. 64(3):1155–1170. ISSN 0025-1909. doi: 10.1287/mnsc.2016.2643.

[3] Jonas C. Ditz, Veronika Lazar, Elmar Lichtmeß, Carola Plesch, Matthias Heck, Kevin Baum, and Markus Langer. Secure human oversight of AI: Exploring the attack surface of human oversight. URL https://arxiv.org/abs/2509.12290. Version Number: 1.

[4] Moira Donegan. Grok is undressing women and children. Don't expect the US to take action, January 2026. URL https://www.theguardian.com/commentisfree/2026/jan/09/grok-undressing-women-children-us-action. Accessed: 2026-01-16.

[5] Thomas Dratsch, Xue Chen, Mohammad Rezazade Mehrizi, Roman Kloeckner, Aline Mähringer-Kunz, Michael Püsken, Bettina Baeßler, Stephanie Sauer, David Maintz, and Daniel Pinto Dos Santos. Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. 307(4):e222176. ISSN 0033-8419, 1527-1315. doi: 10.1148/radiol.222176. URL http://pubs.rsna.org/doi/10.1148/radiol.222176.

[6] Lena Enqvist. 'Human oversight' in the EU Artificial Intelligence act: What, when and by whom? 15(2):508–535. ISSN 1757-9961, 1757-997X. doi: 10.1080/17579961.2023.2245683. URL https://www.tandfonline.com/doi/full/10.1080/17579961.2023.2245683.

[7] European Commission. Directorate General for Communications Networks, Content and Technology. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment.* Publications Office. URL https://data.europa.eu/doi/10.2759/002360.

[8] Michael Feffer, Anusha Sinha, Wesley H. Deng, Zachary C. Lipton, and Hoda Heidari. Red-teaming for generative AI: Silver bullet or security theater? 7:421–437. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31647. URL https://ojs.aaai.org/index.php/AIES/article/view/31647.

[9] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. URL https://arxiv.org/abs/2209.07858. Version Number: 2.

[10] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. 19(1):121–127. ISSN 1067-5027, 1527-974X. doi: 10.1136/amiajnl-2011-000089. URL https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000089.

[11] Ben Green. The flaws of policies requiring human oversight of government algorithms. 45:105681. ISSN 2212473X. doi: 10.1016/j.clsr.2022.105681. URL https://linkinghub.elsevier.com/retrieve/pii/S0267364922000292.

[12] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. 3:1–24. ISSN 2573-0142. doi: 10.1145/3359152. URL https://dl.acm.org/doi/10.1145/3359152.

[13] The Guardian. ChatGPT encouraged Adam Raine's suicidal thoughts. His family's lawyer says OpenAI knew it was broken, August 2025. URL https://www.theguardian.com/us-news/2025/aug/29/chatgpt-suicide-openai-sam-altman-adam-raine. Accessed: 2026-01-16.

[14] Andreas Holzinger, Kurt Zatloukal, and Heimo Müller. Is human oversight to AI systems still possible? 85:59–62. ISSN 18716784. doi: 10.1016/j.nbt.2024.12.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1871678424005636.

[15] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. 11(1):108. ISSN 2158-3188. doi: 10.1038/s41398-021-01224-x. URL https://www.nature.com/articles/s41398-021-01224-x.

[16] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference*. URL https://aisel.aisnet.org/ecis2020_rp/168?utm_source=aisel.aisnet.org%2Fecis2020_rp%2F168&utm_medium=PDF&utm_campaign=PDFCoverPages.

[17] Jacqueline N. Lane, Léonard Boussioux, Charles Ayoubi, Ying Hao Chen, Camilla Lin, Rebecca Spens, Pooja Wagh, and Pei-Hsin Wang. Narrative AI and the human-AI oversight paradox in evaluating early-stage innovations. *Harvard Business School Working Paper*. URL https://aisel.aisnet.org/icis2025/conf_theme/conf_theme/7/.

[18] Johann Laux. Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the european union AI act. 39(6):2853–2866. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-023-01777-z. URL https://link.springer.com/10.1007/s00146-023-01777-z.

[19] Johann Laux and Hannah Ruschemeier. Automation bias in the AI act: On the legal implications of attempting to de-bias human oversight of AI. pages 1–16. ISSN 1867-299X, 2190-8249. doi: 10.1017/err.2025.10033. URL https://www.cambridge.org/core/product/identifier/S1867299X25100330/type/journal_article.

[20] Angelica Lermann Henestrosa and Joachim Kimmerle. Understanding and perception of automated text generation among the public: Two surveys with representative samples in germany. 14(5): Article 353. ISSN 2076-328X. doi: 10.3390/bs14050353. URL https://www.mdpi.com/2076-328X/14/5/353.

[21] Angelica Lermann Henestrosa, Hannah Greving, and Joachim Kimmerle. Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. 138:Article 107445. ISSN 07475632. doi: 10.1016/j.chb.2022.107445. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563222002679.

[22] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. 151:90–103. ISSN 07495978. doi: 10.1016/j.obhdp.2018.12.005.

[23] Margit. E. Oswald and Stefan Grosjean. Confirmation bias. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, reprint edition. ISBN 978-1-84169-351-4.

[24] Raja Parasuraman and Dietrich H. Manzey. Complacency and bias in human use of automation: An attentional integration. 52(3):381–410. ISSN 0018-7208, 1547-8181. doi: 10.1177/0018720810376055. URL https://journals.sagepub.com/doi/10.1177/0018720810376055.

[25] Riana Pfefferkorn. Grok is undressing people online. Here's how to fix it. *The New York Times*, January 2026. URL https://www.nytimes.com/2026/01/12/opinion/grok-digital-undressing.html. Accessed: 2026-01-16.

[26] Gabriel Recchia, Chatrik Singh Mangat, Jinu Nyachhyon, Mridul Sharma, Callum Canavan, Dylan Epstein-Gross, and Muhammed Abdulbari. Confirmation bias: A challenge for scalable oversight. URL https://arxiv.org/abs/2507.19486. Version Number: 1.

[27] Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. Does automation bias decision-making? 51(5):991–1006. ISSN 10715819. doi: 10.1006/ijhc.1999.0252. URL https://linkinghub.elsevier.com/retrieve/pii/S1071581999902525.

[28] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, pages 2495–2507. ACM. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659051. URL https://dl.acm.org/doi/10.1145/3630106.3659051.

[29] The New York Times. Google and character.ai to settle lawsuit over teenager's death.

The New York Times, January 2026. URL https://www.nytimes.com/2026/01/07/technology/google-characterai-teenager-lawsuit.html. Accessed: 2026-01-16.

[30] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. 8(12):2293–2303. ISSN 2397-3374. doi: 10.1038/s41562-024-02024-1. URL https://www.nature.com/articles/s41562-024-02024-1.

[31] Ben Wagner. Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. 11(1):104–122. ISSN 1944-2866, 1944-2866. doi: 10.1002/poi3.198. URL https://onlinelibrary.wiley.com/doi/10.1002/poi3.198.

[32] Sai Wang and Guanxiong Huang. The impact of machine authorship on news audience perceptions: A meta-analysis of experimental studies. 51(7):815–842. ISSN 0093-6502, 1552-3810. doi: 10.1177/00936502241229794. URL https://journals.sagepub.com/doi/10.1177/00936502241229794.

[33] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. Algorithmic decision-making and the control problem. 29(4):555–578. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-019-09513-7. URL http://link.springer.com/10.1007/s11023-019-09513-7.