How to Test for Compliance with Human Oversight Requirements in AI Regulation?

Markus Langer
University of Freiburg, Department of Psychology
Freiburg im Breisgau, Germany

Veronika Lazar
German Federal Office for Information Security
Saarbrücken, Germany

Kevin Baum
German Research Center for Artificial Intelligence, Responsible AI and Machine Ethics
Saarbrücken, Germany

Abstract

Human oversight requirements are a core component of the European AI Act and in AI governance. In this paper, we highlight key challenges in testing for compliance with these requirements. A key difficulty lies in balancing simple, but potentially ineffective checklist-based approaches with resource-intensive empirical testing in diverse contexts where humans oversee AI systems. Additionally, the absence of easily operationalizable standards and the context-dependent nature of human oversight further complicate compliance testing. We argue that these challenges illustrate broader challenges in the future of sociotechnical AI governance.

Keywords

Human Oversight, Auditing, AI Act, Regulation

ACM Reference Format:
Markus Langer, Veronika Lazar, and Kevin Baum. 2025. How to Test for Compliance with Human Oversight Requirements in AI Regulation?. In . ACM, New York, NY, USA, 4 pages.

1 Introduction

Testing for compliance with emerging legislation regarding Artificial Intelligence (AI) such as the European AI Act will be a major task for developers and deployers of AI-based systems when these systems are used for high-risk tasks [7, 12]. Some aspects of this compliance testing will resemble traditional auditing processes for classical software systems and other technologies governed by product safety regulations [7]. For instance, verifying whether AI systems have adequate documentation, ensuring cybersecurity, testing for data protection, and evaluating the accuracy of system outputs could all be achieved by defining standards and quality thresholds. Eventually, compliance testing may also draw on established best practices, such as checklist-based approaches to assess whether implemented processes and technologies comply with standards set by norming bodies.

However, the AI Act and other emerging AI regulations introduce requirements that traditional compliance testing methods cannot easily address. A key example is the requirement for effective human oversight in high-risk AI systems, as outlined in Article 14 of the AI Act [5, 8]. Various countries, including Argentina, Bahrain, Uganda, and South Africa, already enforce similar but less specific requirements for human involvement in AI-driven decision-making [5]. Effective human oversight in the sense of the AI Act will include sub-requirements, such as ensuring that human oversight personnel remain aware of their tendency to over-rely on outputs produced by a high-risk AI system (automation bias, [14]) and that they properly understand the relevant capacities and limitations of the high-risk AI system to adequately monitor its operation (see Article 14 AI Act). Defining clear standards for when oversight is truly “effective,” when overseers recognize automation bias, or when they sufficiently understands AI system limitations remains a challenge [10]. Likewise, developing processes to test for compliance with these requirements will be complex. Although inspiration could be drawn from other areas where human oversight is required (e.g., the European General Data Protection Regulation GDPR), research suggests that human oversight requirements are particularly difficult to operationalize and test (e.g., [15]).

We believe that human oversight requirements are a primary example of the sociotechnical aspects of AI governance. In this position paper, we explore how compliance testing for these requirements may evolve, ranging from simple checklist-based approaches to resource-intensive empirical evaluations of oversight effectiveness in high-risk AI applications. Attending the *Sociotechnical AI Governance* Workshop at CHI’25 will be a great opportunity, as it will allow us to gain new and global insights into sociotechnical AI governance. Additionally, it will provide the intended presenter (Markus Langer) with the opportunity to share his perspective on AI governance in general and human oversight requirements in particular, shaped by his background as a psychology researcher who collaborates closely with experts in law, philosophy, and computer science to enhance our understanding of effective AI governance (see e.g., in our prior work [2, 3, 8, 16])

2 The Possible Future of Testing for Compliance with Human Oversight Requirements

The European AI Act establishes requirements for AI systems classified as "high-risk," including those used in education, public administration, hiring, credit scoring, and medicine. One such requirement is human oversight, as specified in Article 14 (see Appendix A for the full text). It states that "human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular where such risks persist despite the application of other requirements set out in this Section." These other requirements, detailed in Articles 9-13 and Article 15 of the AI Act, cover risk management, data governance, technical documentation, record keeping, transparency, accuracy, robustness, and cybersecurity.

Conference’17, July 2017, Washington, DC, USA
2025. ACM ISBN 978-1-4503-XXXX-X/18/06

Some key requirements of Article 14 include that human oversight personnel should be able to understand the capacities and limitations of AI systems and correctly interpret outputs. They should remain aware of their tendency for automation bias (which, according to the AI Act, refers to overtrust in AI outputs), decide when not to use AI outputs, and override decisions when necessary. They are also supposed to intervene or interrupt a system – for example, using a stop button or a similar mechanism to halt the system in a safe state.

Standards and norms are currently being developed to guide compliance testing for human oversight requirements, including the Trustworthiness Framework developed by CEN/CENELEC and the ISO/IEC CD 42105. Part of these standards and norms will be informed by Article 14 of the AI Act and broader international governance perspectives on human oversight. While still in development, we anticipate a future of compliance testing for human oversight along a continuum between simple, checklist-based approaches and empirical testing of the effectiveness of human oversight in specific contexts.

A checklist approach would follow the model of existing compliance testing methods [7, 15] translating Article 14 requirements into assessable items for internal or external auditors. A checklist might include items such as: "The human oversight person has been made aware of their tendency to overtrust outputs of AI-based systems", "The human oversight person has received adequate training that enables them to understand the capacities and limitations of the AI-based system they oversee", or "There is a stop button that allows the human oversight person to intervene in the operation of the AI-based system". Clearly, this list is not exhaustive, and the requirements would need to be refined. While such checklists could provide a straightforward compliance mechanism, they may fall short of the AI Act's broader goal of effectively mitigating AI-related risks [7]. Moreover, developing an exhaustive checklist will be challenging. The examples above are direct translations from Article 14. Clearly there can be an infinite number of requirements relating to an infinite number of degrees of specificity, for instance, requirements regarding the person who will be the human oversight person (e.g., specific skills they need to possess), or work design of human oversight jobs (e.g., specific maximum durations for human oversight tasks) [16]. It also remains unclear how frequently oversight processes should be reevaluated or whether checklist-based assessments should be triggered by evidence of non-compliance.

Empirical testing of the effectiveness of human oversight in specific contexts could address the limitations of checklists. This approach would require testing the actual effectiveness of human oversight in high-risk contexts and empirically demonstrating compliance with AI regulatory requirements [5]. It could involve studies where human oversight personnel monitor AI systems for a set duration, assessing whether they detect erroneous or problematic outputs, intervene in system operation when necessary, and accurately override inadequate AI-generated decisions.

Another option could involve comparing different human oversight designs to determine which best meets regulatory requirements. Human oversight design, as outlined by Sterz et al. [16] is a sociotechnical design question. It encompasses technical aspects (e.g., optimizing explainability approaches to support human oversight), individual factors (e.g., selecting and training oversight personnel), and environmental conditions (e.g., job design and working conditions). For instance, a controlled experiment could test various explainability approaches to assess which most effectively supports oversight [1, 6, 9]. The main advantage of this approach is that it provides empirical evidence on the effectiveness of human oversight and how to optimize it.

However, this approach demands significant resources for planning, conducting, analyzing, and interpreting studies. Empirical testing for the effectiveness of human oversight requires substantial expertise with empirical methods and user studies – researchers and practitioners, for instance, with a background in human-computer interaction or psychology will be required to adequately conduct empirical testing, interpreting the results, and providing recommendations on how to optimize human oversight design. Moreover, transferring insights across contexts may be challenging, as oversight effectiveness and requirements can be context-dependent. Determining the frequency of empirical testing—whether after AI system updates or personnel changes—adds further complexity. Furthermore, deriving reliable conclusions often requires multiple studies (e.g., on the effects of different explainability approaches), suggesting that oversight requirements may need to be informed by high-quality meta-analyses that synthesize findings across studies for broader applicability.

Checklist-based approaches and empirical approaches are clearly not the only possible options of testing for compliance with human oversight requirements, but they illustrate two contrasting futures: checklists offer efficiency but risk assuming compliance with oversight requirements without ensuring oversight effectiveness, while empirical approaches provide validation of effectiveness but may be too resource-intensive and context-dependent. One obvious solution would be to combine the advantages of both approaches. For instance, if a checklist approach does not lead to sufficiently clear results, this could call for an empirical approach to augment it.

However, both approaches face additional challenges. First, some regulatory requirements may be difficult to translate into testable standards due to the lack of a clear ground truth. For instance, AI regulations aim to mitigate risks to fundamental rights, such as fairness, but fairness itself is complex, with multiple definitions that cannot all be satisfied simultaneously [8]. Additionally, some requirements involve psychological factors that are challenging to operationalize, such as ensuring human oversight personnel understand AI limitations or assessing automation bias [10, 14]. Prior work (e.g., [15]) has, for example, highlighted that compliance testing struggles with fairness requirements and aspects related to transparency and explainability. Addressing these challenges may require expertise in psychology and empirical research to develop meaningful standards and effective compliance testing methods.

Without concrete and testable standards, uncertainty will persist for providers and deployers of AI regarding their legal compliance – an issue that may be particularly pronounced for small businesses that lack the resources to establish an AI compliance department [15]. This uncertainty could lead to situations where the implementation of AI-based systems will be hampered in such businesses. At the same time, without concrete standards, virtually any implementation could be considered compliant [4]. This issue is especially problematic when audits rely on post-hoc rationalizations of human oversight implementations. In hindsight, any approach to human oversight could be justified as the "best possible" option.

Second, human oversight requirements may be context-dependent. For instance, the level of oversight needed may vary depending on the risk associated with the application context. Stricter oversight requirements may apply to AI use in the public sector compared to the private sector. Additionally, the required skills, expertise, and tasks of human oversight personnel can also vary significantly [15]. In real-time contexts, such as autonomous vehicles, sustained vigilance over long periods may be necessary, whereas in areas like hiring and credit scoring, human oversight personnel may require training in ethical and moral reasoning.

3 Concluding Thoughts and Next Steps

The challenges of testing for compliance with human oversight requirements reflect broader difficulties in sociotechnical AI governance. As AI governance shifts from ensuring "good" technological products to "good" sociotechnical systems, defining standards will be complex, particularly when psychological concepts are involved. A key example is emotion recognition systems [13]. According to the Act, AI systems that automatically infer emotions (e.g., sadness) in high-risk contexts are prohibited but it is allowed to infer physical states (e.g., fatigue). This raises questions such as: Is fatigue a purely physical state from lack of sleep or a symptom of depression? If linked to depression, would its detection be permitted? Beyond the challenges posed by psychological concepts in AI regulation, additional difficulties arise when AI governance seeks to mitigate risks for which no clear ground truth exists (e.g., risks of discrimination [15]) or when it remains uncertain whether risks have been effectively mitigated. For instance, was a fairness monitoring tool truly successful if it detects only one specific type of fairness violation in AI outputs [3, 8]?

The next steps in testing for compliance with AI regulation are currently being developed. Standardization and norming bodies are working to operationalize the requirements outlined in AI regulation. We anticipate that emerging standards and norms will fuel the debate on how to effectively test for compliance. The challenges outlined in this article will continue to require input from researchers, practitioners, and policymakers to ensure that AI governance effectively reduces the risks associated with AI systems while enhancing safety in their implementation – without placing an undue burden on providers and deployers through resource-intensive, context-dependent compliance testing.

In the case of human oversight, the key challenges for the near future are to (a) establish a middle ground or a feasible combination between checklists and empirical testing, (b) develop standards and norms that are informed by and adapt to the latest research in HCI, psychology, and related fields [11] – such as methods for preventing automation bias or effectively preparing and supporting humans to detect inaccurate and problematic outputs, and (c) evaluate the impact of human oversight requirements in AI practice. Finally, we want to highlight the crucial importance of expertise on the human factor in human-AI interaction for designing and testing for the effectiveness of human oversight. As AI governance evolves beyond technological improvement to optimizing sociotechnical systems for high-risk tasks, we believe research(ers) from HCI, psychology, and related fields should play a key role in providing insights on how to optimize the technology, how to design the jobs and environments where humans and AI-based systems interact, and how to prepare and support human oversight personnel.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[2] Kevin Baum, Sebastian Biewer, Holger Hermanns, Sven Hetmank, Markus Langer, Anne Lauber-Rönsberg, and Sarah Sterz. 2024. *Taming the AI Monster: Monitoring of Individual Fairness for Effective Human Oversight.* Springer Nature Switzerland, 3–25. https://doi.org/10.1007/978-3-031-66149-5_1

[3] Sebastian Biewer, Kevin Baum, Sarah Sterz, Holger Hermanns, Sven Hetmank, Markus Langer, Anne Lauber-Rönsberg, and Franz Lehr. 2024. Software doping analysis for human oversight. *Formal Methods in System Design* (April 2024). https://doi.org/10.1007/s10703-024-00445-2

[4] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* ACM, 598–609. https://doi.org/10.1145/3442188.3445921

[5] Ben Green. 2021. The Flaws of Policies Requiring Human Oversight of Government Algorithms. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3921216

[6] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23).* Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[7] Khoa Lam, Benjamin Lange, Borhane Blili-Hamelin, Jovana Davidovic, Shea Brown, and Ali Hasan. 2024. A Framework for Assurance Audits of Algorithmic Systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24).* ACM, 1078–1092. https://doi.org/10.1145/3630106.3658957

[8] Markus Langer, Kevin Baum, and Nadine Schlicker. 2024. Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs. *Minds and Machines* 35, 1 (Nov. 2024). https://doi.org/10.1007/s11023-024-09701-0

[9] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (July 2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[10] Johann Laux. 2023. Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI amp; SOCIETY* 39, 6 (Oct. 2023), 2853–2866. https://doi.org/10.1007/s00146-023-01777-z

[11] Johann Laux and Hannah Ruschemeier. 2025. Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI. arXiv:2502.10036 [cs.CY] https://arxiv.org/abs/2502.10036

[12] Jakob Mökander, Maria Axente, Federico Casolari, and Luciano Floridi. 2022. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines* 32, 2 (Nov. 2022), 241–268. https://doi.org/10.1007/s11023-021-09577-4

[13] Alexandra Prégent. 2025. Is there not an obvious loophole in the AI act's ban on emotion recognition technologies? *AI amp; SOCIETY* (March 2025). https://doi.org/10.1007/s00146-025-02289-8

[14] Hannah Ruschemeier and Lukas Hondrich. 2024. Automation Bias in Public Administration - an Interdisciplinary Perspective From Law and Psychology. *SSRN Electronic Journal* (2024). https://doi.org/10.2139/ssrn.4736646

[15] Teresa Scantamburlo, Paolo Falcarin, Alberto Veneri, Alessandro Fabris, Chiara Gallese, Valentina Billa, Francesca Rotolo, and Federico Marcuzzi. 2024. Software Systems Compliance with the AI Act: Lessons Learned from an International Challenge. In *Proceedings of the 2nd International Workshop on Responsible AI Engineering (RAIE '24).* ACM, 44–51. https://doi.org/10.1145/3643691.3648589

[16] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24).* ACM, 2495–2507. https://doi.org/10.1145/3630106.3659051

A Appendix

A.1 Content of Article 14 of the European AI Act: Human Oversight

(1) High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.

(2) Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular where such risks persist despite the application of other requirements set out in this Section.

(3) The oversight measures shall be commensurate with the risks, level of autonomy and context of use of the high-risk AI system, and shall be ensured through either one or both of the following types of measures:

(a) measures identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service;

(b) measures identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the deployer.

(4) For the purpose of implementing paragraphs 1, 2 and 3, the high-risk AI system shall be provided to the deployer in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate:

(a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;

(b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;

(c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;

(d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;

(e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.

(5) For high-risk AI systems referred to in point 1(a) of Annex III, the measures referred to in paragraph 3 of this Article shall be such as to ensure that, in addition, no action or decision is taken by the deployer on the basis of the identification resulting from the system unless that identification has been separately verified and confirmed by at least two natural persons with the necessary competence, training and authority.

The requirement for a separate verification by at least two natural persons shall not apply to high-risk AI systems used for the purposes of law enforcement, migration, border control or asylum, where Union or national law considers the application of this requirement to be disproportionate.