New BIOTECHNOLOGY 85 (2025) 59–62

Contents lists available at ScienceDirect

New BIOTECHNOLOGY

journal homepage: www.elsevier.com/locate/nbt

Editorial

Is human oversight to AI systems still possible?

ARTICLE INFO

Keywords:
Artificial intelligence
Biotechnology
Deep learning
Digital transformation
Machine learning

ABSTRACT

The rapid proliferation of artificial intelligence (AI) systems across diverse domains raises critical questions about the feasibility of meaningful human oversight, particularly in high-stakes domains such as new biotechnology. As AI systems grow increasingly complex, opaque, and autonomous, ensuring responsible use becomes a formidable challenge. During our editorial work for the special issue "Artificial Intelligence for Life Sciences", we placed increasing emphasis on the topic of "human oversight". Consequently, in this editorial we briefly discuss the evolving role of human oversight in AI governance, focusing on the practical, technical, and ethical dimensions of maintaining control. It examines how the complexity of contemporary AI architectures, such as large-scale neural networks and generative AI applications, undermine human understanding and decision-making capabilities. Furthermore, it evaluates emerging approaches—such as explainable AI (XAI), human-in-the-loop systems, and regulatory frameworks—that aim to enable oversight while acknowledging their limitations. Through a comprehensive analysis, the picture emerged while complete oversight may no longer be viable in certain contexts, strategic interventions leveraging human-AI collaboration and trustworthy AI design principles can preserve accountability and safety. The discussion highlights the urgent need for interdisciplinary efforts to rethink oversight mechanisms in an era where AI may outpace human comprehension.

Human oversight in AI systems

Human oversight in artificial intelligence (AI) has emerged as a critical mechanism in the governance of AI, tasked with enhancing system accuracy and safety, upholding human values, and fostering trust in the technology [1,2] and became a cornerstone of the European AI Act for application of high risk AI systems (ref). However, empirical evidence highlights significant shortcomings in human oversight. Oversight tasks often suffer from issues such as a lack of information or competence among overseers or harmful incentives that undermine their effectiveness. Addressing these challenges is essential for ensuring robust oversight in AI governance [3].

Biological and biotechnological applications often require interpretable and trustworthy results, as they directly impact human health, environmental safety, and regulatory compliance. The reliance on deep learning models, which are often opaque, and autonomous AI systems that operate without continuous human intervention, complicates the oversight process. Furthermore, the dynamic and uncertain nature of biological data amplifies the risks of errors, biases, and unintended consequences in AI-driven analyses [4,5].

Challenges of human oversight in AI systems

Human oversight in AI systems faces several significant challenges, stemming primarily from the complexity, scale, and partial or full autonomy of modern AI technologies where the promises are high [6]. However, many AI models, especially deep learning systems, operate as "black boxes," making their decision-making processes opaque and difficult for humans to interpret. This lack of transparency complicates oversight, particularly in high-risk applications where understanding the rationale behind AI decisions is critical and may harm humans.

The autonomy of AI systems further exacerbates the issue, as these technologies often function with minimal human intervention, making real-time monitoring difficult. This challenge is heightened by the scale at which AI systems operate, processing vast amounts of data and performing tasks at speeds beyond human cognitive capability, which renders continuous human supervision infeasible.

Data-related issues also play a critical role. Poor data quality, biases in proper representation of relevant features in training datasets, and systemic inequities can compromise AI reliability and ethical alignment. Oversight becomes particularly challenging when these biases are subtle or deeply embedded, requiring sophisticated methods to detect and mitigate them. Furthermore, the ethical and societal implications of AI decisions, especially in domains such as healthcare, access to social systems and criminal justice, demand nuanced oversight mechanisms that are often underdeveloped or even lacking entirely.

Resource constraints add another layer of difficulty. Effective oversight requires proper presentation of information to users as well as substantial human expertise in the interpretation of AI-generated results, computational resources, and time, which are often limited by organizational priorities or financial pressures. This is compounded by gaps in regulatory and policy frameworks, which have struggled to keep pace with the rapid advancement of AI technologies, creating ambiguities in accountability and governance. In this context the European AI

https://doi.org/10.1016/j.nbt.2024.12.003

Available online 13 December 2024

Act provides an important legal framework that places major emphasis on several aspects related to human oversight. Also at UN level human oversight has been promoted in context of transparency, predictability, reliability and understandability of AI in its resolution of March 2024 [7].

There is also a growing tendency to over-rely on automation for oversight, which can diminish the role of human judgment and lead to automation bias, where humans uncritically accept AI outputs. This problem is further compounded by a shortage of professionals with the interdisciplinary knowledge necessary to oversee complex AI systems effectively. Cognitive and psychological limitations also play a role, as the demands of interpreting complex AI outputs and maintaining attention during prolonged monitoring can result in oversight fatigue or errors.

Finally, the dynamic and uncertain environments in which many AI systems operate, such as in biological research or climate modeling, add an additional layer of complexity to oversight. Resistance to stringent oversight measures from developers and stakeholders, who may view such mechanisms as restrictive to innovation, further underscores the difficulty of achieving effective governance. These challenges necessitate innovative solutions that combine technical advances, robust regulatory frameworks, and interdisciplinary collaboration to ensure AI systems remain safe, reliable, and aligned with human values.

The development of human oversight in generative AI

With the breakthrough of large language models (LLMs), the concept of human oversight, formerly often referred to as 'explainable AI' (xAI), has changed fundamentally [8,9]. These models present researchers and developers with immense challenges: The size and complexity of modern language models makes classical explainability methods difficult to apply. At the same time, the increasing dependence on generative AI requires reliable mechanisms to ensure the veracity of statements and avoid hallucinations - the generation of plausible-sounding but false information.

Traditional approaches to explainability, which have been effective in smaller models, face significant limitations when applied to large language models (LLMs). Methods like feature attribution, salience mapping, or attention mechanisms provide partial insights into specific areas of the model but fail to offer a comprehensive understanding of its internal decision-making logic. Gradient-based techniques such as Integrated Gradients or Layer-wise Relevance Propagation (LRP) can highlight influential tokens but do not explain the overarching reasoning of the model. Tools like SHAP or LIME, which approximate local decisions, lose their effectiveness when dealing with the complexity of large-scale text processing. Similarly, while probing methods and Concept Activation Vectors (CAVs) help analyse linguistic patterns, much of the behavior and emergent properties of LLMs remain opaque, highlighting the challenges in fully understanding these sophisticated systems [10].

Human oversight: the human in the loop

Human oversight is increasingly being used to compensate for the weaknesses of LLMs. The following approaches play a central role here:

- **Human-in-the-loop (HITL):** humans evaluate and correct model responses during training or in real-time operation. The method of reinforcement learning with human feedback (RLHF), in which human feedback flows directly into the optimisation of the model, is particularly effective.
- **Audit and validation**: Interactive debugging tools enable users to scrutinise model responses in a targeted manner. Benchmark sets help to ensure fair and comprehensible decisions.
- **Rule-based interventions**: In sensitive applications such as medicine, rule-based systems can filter the outputs of LLMs and forward them for human review.
- **Explainable user interfaces:** Transparent interfaces that visualise uncertainties and alternative answers improve trust in AI-powered systems.
- **Hybrid approaches**: The combination of expert systems and LLMs limits the autonomy of the models and enables continuous human validation.

The analysis of large language models (LLMs) can draw inspiration from brain research, where techniques like transcranial magnetic stimulation (TMS) or optogenetics are used to selectively deactivate neuronal regions to study their functions. Similarly, in LLMs, ablation analysis involves deactivating specific components, such as layers or attention heads, to examine their roles in processing inputs, while interference through noise introduces experimental disruptions to model activations to identify robust and sensitive elements. These methods enhance our understanding of the modularity and emergent properties of LLMs, paralleling how brain research uncovers the functional organisation of neural systems.

One of the biggest problems with LLMs is the tendency to hallucinate. These occur when models generate information that is not anchored in their training data or inputs. Approaches to minimise this problem include:

- **Fact-based training:** using curated and regularly updated data sources reduces bias and error.
- **Knowledge base integration**: LLMs can access external data sources such as Wikidata to validate answers.
- **Retrieval Augmented Generation**. The results of LLMs are restricted to information form specified and quality controlled documents (e.g., handbooks of learned societies) that markedly facilitates interpretation of results by the user.
- **Post-hoc fact checking**: Automated fact checkers verify the accuracy of model answers before they are presented.
- **Self-monitoring**: Models can be trained to signal uncertainties or formulate their answers more cautiously.
- **User feedback**: Mechanisms such as RLHF utilise user feedback to correct erroneous or hallucinated patterns.

The further development of human oversight and explainability approaches is crucial in order to utilise the potential of large language models responsibly. The interdisciplinary approach - inspired by methods from brain research and combined with modern validation mechanisms - offers a promising perspective. Nevertheless, the challenge remains to make AI systems not only more powerful, but also more transparent, trustworthy and secure. The integration of human expertise will play a central role in this.

Regulatory and ethical considerations

Regulatory and ethical considerations are central to the implementation of human oversight in AI systems, particularly as these systems grow more complex and autonomous. A key regulatory challenge lies in the rapid pace of AI development, which often outstrips the creation of robust legal frameworks. This lag results in significant gaps in accountability, particularly in determining liability for decisions made by AI systems. Governments and international organizations must balance fostering innovation with implementing regulations that ensure safety, fairness, and transparency.

From an ethical perspective, oversight mechanisms must promote safe, secure and trustworthy AI by addressing issues such as AI being human-centric, reliable, explainable, ethical, inclusive, in full respect, promotion and protection of human rights and international law, privacy preserving, sustainable development oriented, and responsible

[11]. In this context, the use of biased or incomplete training data can lead to systemic inequities, amplifying societal disparities rather than mitigating them. Human oversight must therefore include rigorous evaluation of data quality and fairness metrics, alongside ethical guidelines tailored to specific application domains.

Privacy concerns also feature prominently in regulatory and ethical debates. AI systems often rely on vast amounts of data, including sensitive personal or biological information. Oversight must ensure that data is collected, processed, and stored in compliance with privacy laws and ethical principles, while balancing the trade-off between data utility and individual rights.

Another critical consideration is the explainability and interpretability of AI systems. Regulations increasingly demand that AI outputs be understandable to humans, particularly in high-stakes areas like healthcare, finance, and criminal justice. This is also an essential requirement where people using AI systems are fully accountable for their decision making. Ensuring transparency is both a technical and ethical imperative, as it allows stakeholders to validate AI decisions and fosters trust in these systems.

Ethical considerations also encompass the broader societal impacts of AI. Human oversight must ensure that AI systems align with societal values, avoiding harm and promoting well-being. This requires interdisciplinary collaboration between technologists, ethicists, policymakers, and domain experts to establish context-specific oversight mechanisms that address the potential for misuse or unintended consequences.

Regulatory frameworks must also consider the global nature of AI development and deployment [12]. International cooperation is necessary to harmonize standards and prevent regulatory arbitrage, where companies exploit less stringent oversight in certain jurisdictions. Furthermore, equitable access to AI technologies must be ensured, preventing a concentration of benefits in a few regions or demographics while others are left behind.

Ultimately, regulatory and ethical considerations demand a proactive approach to human oversight, emphasizing accountability, transparency, fairness, and inclusivity. These principles must be embedded not only in governance policies but also in the design and implementation of AI systems, ensuring that their development serves the broader interests of society.

Future directions and recommendations

Future directions for human oversight in AI systems emphasize the need for interdisciplinary approaches, technical innovation, and policy reform to address the growing complexity and impact of AI technologies. One critical recommendation is to integrate explainability and causability [13] into AI systems from the design phase. By embedding transparency as a core feature, developers can facilitate human understanding and validation of AI decision-making processes. Coupled with advancements in explainable AI (XAI), these efforts can help reduce the opacity of complex models and empower oversight mechanisms.

Another key direction involves enhancing human-AI collaboration through the development of human-in-the-loop frameworks and new types of human-AI interfaces [14]. These systems allow humans to interact with and influence AI behavior in real-time, striking a balance between autonomy and control. This approach is particularly relevant in high-risk domains where human expertise and contextual judgment remain indispensable. Training programs and educational initiatives are also essential to equip stakeholders with the skills necessary to oversee AI systems effectively. Interdisciplinary training that combines technical, ethical, and domain-specific knowledge will prepare professionals to navigate the challenges posed by increasingly sophisticated AI technologies.

Policy reforms must address existing gaps at legal, regulatory and standards levels by establishing clear accountability structures for AI systems. This should include internationally accepted and aligned governance models, standards and guidelines on data usage, bias detection, and fairness to ensure ethical alignment. International cooperation will be crucial to harmonize standards and prevent regulatory fragmentation, particularly as AI systems often operate across borders. New frameworks should also prioritize adaptive governance models that can evolve in response to emerging AI capabilities and challenges.

Future research should focus on developing scalable oversight tools that can monitor AI systems operating at unprecedented speeds and volumes. Innovations such as AI-assisted monitoring and validation mechanisms can augment human oversight, enabling more effective management of large-scale deployments. Additionally, fostering collaboration between academia, industry, and policymakers will accelerate the development of robust oversight solutions tailored to diverse applications and societal needs.

A forward-looking approach must also emphasize equity and inclusivity, ensuring that AI oversight mechanisms do not disproportionately burden marginalized communities or exacerbate existing inequalities. Ethical considerations should guide the development and deployment of oversight tools, ensuring that they promote societal well-being and mitigate harm. Ultimately, the future of human oversight in AI systems depends on fostering trust, accountability, and transparency while enabling the responsible innovation needed to address complex global challenges.

Credit authorship contribution statement

All authors contributed to conceptualization, writing, review and editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. KZ is co-founder and CEO of Zatloukal Innovations GmbH. This work does not raise any ethical issues.

Acknowledgements

Parts of this work have received funding from the Austrian Science Fund (FWF), Project P-32554 (Explainable Artificial Intelligence) and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874662 (HEAP), and No. 101079183 (BioMedAI). This publication reflects only the authors' view, and the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] Dietterich TG, Horvitz EJ. Rise of concerns about AI: reflections and directions. Commun ACM 2015;58(10):38–40. https://doi.org/10.1145/2770869.
[2] Green B. The flaws of policies requiring human oversight of government algorithms. Comput Law Secur Rev 2022;45:105681. https://doi.org/10.1016/j.clsr.2022.105681.
[3] Laux J. Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. AI Soc 2023:1–14. https://doi.org/10.1007/s00146-023-01777-z.
[4] Shneiderman B. Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Hum–Comput Interact 2020;36(6):495–504. https://doi.org/10.1080/10447318.2020.1741118.
[5] Schmid U. Trustworthy artificial intelligence: comprehensible, transparent and correctable. In: Werthner H, Ghezzi C, Kramer J, Nida-Rümelin J, Nuseibeh B, Prem E, editors. Introduction to Digital Humanism. Cham: Springer; 2024. p. 151–64. https://doi.org/10.1007/978-3-031-45304-5_10.
[6] Saenz AD, Harned Z, Banerjee O, Abràmoff MD, Rajpurkar P. Autonomous AI systems in the face of liability, regulations and costs. NPJ Digit Med 2023;6(1):185. https://doi.org/10.1038/s41746-023-00929-1.
[7] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU,

(EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024.
[8] Retzlaff CO, Angerschmid A, Saranti A, et al. Post-Hoc vs Ante-Hoc explanations: XAI design guidelines for data scientists. Cogn Syst Res 2024;86(8):101243. https://doi.org/10.1016/j.cogsys.2024.101243.
[9] Combi C, Amico B, Bellazzi R, et al. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;133(11):102423. https://doi.org/10.1016/j.artmed.2022.102423.
[10] Kraišniković C, Harb R, Plass M, Al Zoughbi W, Holzinger A, Müller H. Fine-tuning language model embeddings to reveal domain knowledge: an explainable artificial intelligence perspective on medical decision making. Eng Appl Artif Intell 2025; 139:109561. https://doi.org/10.1016/j.engappai.2024.109561.
[11] United Nations Resolution A/78/L.49 of 11 March 2024, 2024.
[12] Müller H, Holzinger A, Plass M, Brcic L, Stumptner C, Zatloukal K. Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. N Biotechnol 2022;70: 67–72. https://doi.org/10.1016/j.nbt.2022.05.002.
[13] Plass M, Kargl M, Kiehl TR, et al. Explainability and causability in digital pathology. J Pathol Clin Res 2023;9(4):251–60. https://doi.org/10.1002/cjp2.322.
[14] Holzinger A, Kargl M, Kipperer B, Regitnig P, Plass M, Müller H. Personas for artificial intelligence (AI) an open source toolbox. IEEE Access 2022;10:23732–47. https://doi.org/10.1109/ACCESS.2022.3154776.

Andreas Holzinger*
Human-Centered AI Lab, Institute of Forest Engineering, Department for Ecosystem Management, Climate and Biodiversity, University of Natural Resources and Life Sciences Vienna, Austria
Information Science and Machine Learning Group, Diagnostic and Research Institute of Pathology, Medical University Graz, Austria

Kurt Zatloukal, Heimo Müller
Information Science and Machine Learning Group, Diagnostic and Research Institute of Pathology, Medical University Graz, Austria

* Corresponding author at: Human-Centered AI Lab, Institute of Forest Engineering, Department for Ecosystem Management, Climate and Biodiversity, University of Natural Resources and Life Sciences Vienna, Austria.
E-mail address: andreas.holzinger@boku.ac.at (A. Holzinger).