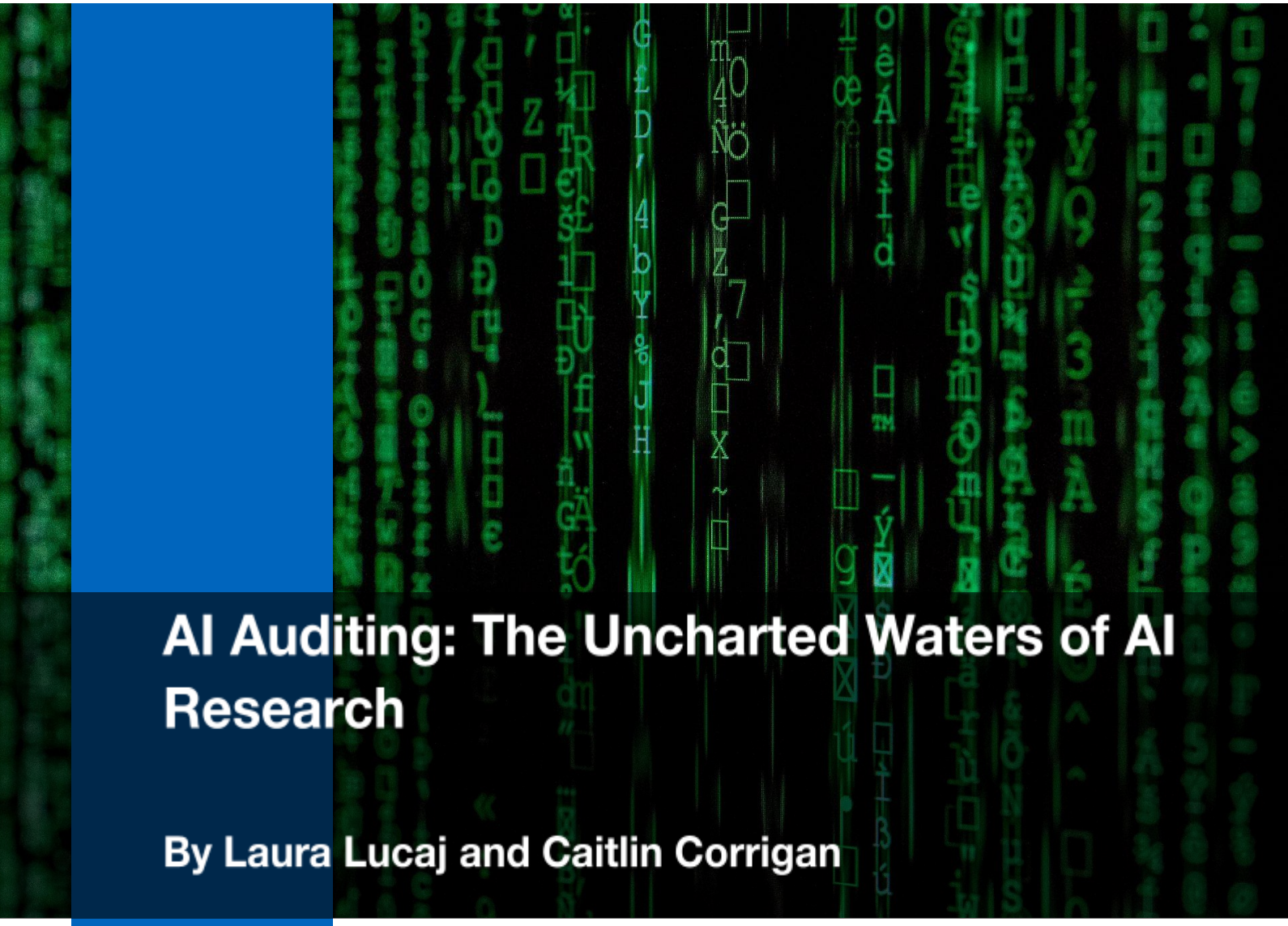
IEAI
Technical University of Munich
School of Social Science and Technology
Institute for Ethics in Artificial Intelligence

Research Brief – August 2022



AI Auditing: The Uncharted Waters of AI Research

By Laura Lucaj and Caitlin Corrigan

Auditing of AI systems has gained considerable recognition as an opportunity to harness the potential of AI models, as well as for detecting and mitigating the problematic patterns and consequences of their deployment in sensitive contexts such as healthcare, hiring or mobility, to name a few. In the last few years, research has shown how AI systems can produce predictions that disproportionately affect vulnerable minorities and contribute to the reinforcement of structural biases that already exist in societal contexts. Hence, algorithmic AI auditing procedures are increasingly necessary to understand, detect and mitigate the potential unintended consequences of the deployment of AI-enabled technology. This research brief addresses the topic of AI auditing and the need for an interdisciplinary conversation on the topic.

AI systems hold the promise of helping humans address many challenges, such as reducing biodiversity loss and managing climate change, increasing efficiency of processes, enabling smart traffic control and sustainable mobility development, and improving healthcare diagnostics (Eastwood et al, 2021; Cheong et al., 2022; Manyika & Sneader, 2018; Englund et al, 2021; Jiang et al., 2017). However, ethical challenges such as disproportionately affecting vulnerable minorities (e.g., discriminating female candidates in the context of hiring through AI-enabled recruiting systems) continue to emerge in this space, while corresponding accountability and oversight mechanisms are lacking (Yam & Skorburg, 2021). This absence of oversight enables the un-scrutinized deployment of such systems in sensitive contexts such as criminal justice, hiring or healthcare, among others (McKay, 2020; Panch et al., 2019; Dattner et al, 2019). Hence, AI auditing procedures are increasingly necessary to understand, detect and mitigate the potential unintended consequences of the deployment of AI-enabled technology. This research brief highlights the increasing importance of addressing AI auditing and contends that it is key to do so in an interdisciplinary setting with collaboration between policymakers, academia, and AI developers.

The Need for AI Auditing and Impact Assessment Practices

Audit procedures for AI systems provide an overview of the system's present and past performance and enable monitors to preemptively address, manage and mitigate potential risks (Falco et al., 2021; Brundage et al., 2020). Audits can be performed internally by the organization developing the technology or by an external entity such as third-party auditing companies (Knowles & Richards, 2021; Raji et al., 2020).

Surprisingly, in spite of the increased awareness of auditing being a fundamental process to address challenges that AI systems can bring along, little work has been conducted so far in addressing the operational challenges of such complex and interdisciplinary research fields. Few researchers have provided guidance on how to address issues of auditing AI in practice throughout the entire lifecycle of the system. Thus, effective, actionable, and comprehensive methodologies to understand, address and mitigate such impacts remain under-investigated.

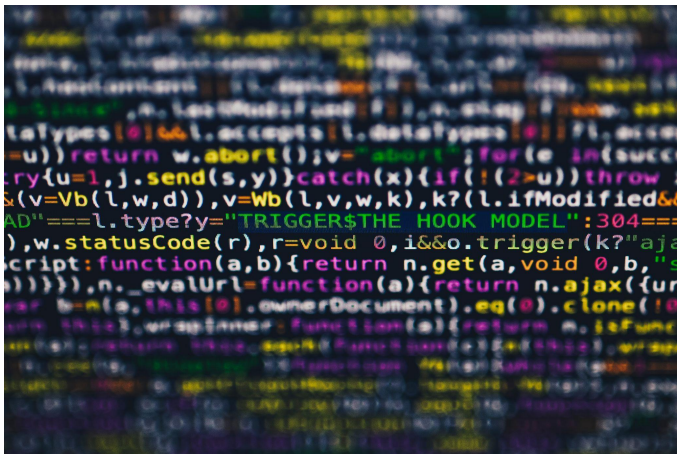
In the past few years, numerous guidelines around the deployment of AI have been published by industrial, academic, non-profit, as well as governmental institutions (Brundage et al., 2020; Hagendorff, 2020; Falco et al., 2021). Two leading standard bodies, such as the IEEE and the ISO have been publishing standards for AI systems to improve market efficiency and address some of the ethical concerns (Chihon, 2019). The standards address a multitude of practices, such as frameworks for AI systems using machine learning, and algorithmic bias considerations, to name a few (ISO 2018, IEEE, 2022). In spite of such standards providing a path towards a global solution, they are not sufficient in achieving a proper mechanism of governance oversight that enables the right checks and balances.

AI auditing procedures are increasingly necessary to understand, detect and mitigate the potential unintended consequences of the deployment of AI-enabled technology.

The European Commission (EC), has been leading worldwide with regulatory frameworks for developing human-centered AI systems. For instance, the Ethics Guidelines for Trustworthy AI, proposed in 2019, as well as the White Paper on AI in 2020 (European Commission, 2019; European Commission, 2020). In 2021, the EU

Artificial Intelligence Act was issued as a proposal for legislating AI systems (European Commission, 2021) with the goal of guaranteeing the development of safe AI systems that conform with the human-centric values of the European Union. Recently, the EC has worked on translating the principles into actionable practices through tools such as the Assessment List for Trustworthy Artificial Intelligence (European Commission, 2020), whose aim is to help developers to assess the impact of their AI systems.

However, practitioners are still increasingly struggling to translate guidelines into their decision-making processes (Hagendorff, 2020). Assurance requirements for organizations are non-trivial (Falco et al, 2021). All these initiatives are still lacking operational guidance for practitioners to take effective, actionable measures to effectively assess the impact of the AI systems they develop throughout the lifecycle.



Understanding potential practices in auditing

Auditing has been a standard practice in numerous domains and industries already relying heavily on standardized practices to assess the quality and security of the systems developed, such as finance, aerospace, or healthcare (Raji et al., 2020; Shneidermann, 2020). In the last few years, several practices have been explored by researchers, who investigated best practices for auditing AI systems. For instance, practices that have been adopted in other domains, such as **audit trails, verification, and bias testing**, as well as **explainable user interfaces,** could all be important aspects to consider (Shneiderman, 2020). Such measures enable to log the necessary information on the practices conducted to assess the impact of the system, as well as transparently communicating to the users how the system operates and the safe extent of its deployment (Shneiderman, 2020). Others have examined templates to enable **documentation practices**, essential for creating logs of information in order to understand the choices that were made in the development phase and understand their potential impact at a later deployment phase (Arnold et al., 2019; Richards, 2020). Similar to the supplier's declaration of conformity (SDoC), which provides information on how a product conforms to the technical standards or regulations enforced in the country it is deployed, **FactSheets** have been proposed as well (Arnold et al., 2019). They constitute a comprehensive documentation framework, aiming to record the practices deployed in the development of the AI system and disclose the intended purpose for deployment (Arnold et al., 2019; Richards, 2020). Such practices could significantly contribute to an increase in consumers' trust in AI-enabled products and would enable monitoring of potential ethical concerns emerging in the design and development phase (Arnold et al., 2019; Richards, 2020). Documentation practices constitute essential steps that should be embedded into auditing procedures. However, on their own, they are not sufficient to address all the challenges of the impact assessment of AI-enabled systems.

Different **auditing tools** for specific ethical requirements such as **fairness and explainability** have been explored, and mitigation measures for adhering to the ethical standards have been proposed (Koshiyama et al., 2021). However, a general procedure for bringing consistency to all the phases of the components of an AI model is missing.

The comprehensive framework for internal auditing of AI systems proposed by Raji et al. (2020) has been a major contribution to the field by enabling a proactive set of interventions throughout the lifecycle of a system. The framework tackles the practices necessary along the pipeline of an AI system to record important design decisions and to identify the causal relation between such decisions and the risks that might emerge and relate to ethical failures (Raji et al., 2020). The authors propose five different stages that are **Scoping, Mapping, Artifact Collection, Testing, and Reflection,** all-encompassing the necessary documentation requirements and enabling closing of the accountability gap for each fundamental process of the analysis of an AI system (Raji et al., 2020). This process contributes significantly to the identification of which documentation practices can contribute to the evaluation of an AI system's work in the product development lifecycle. Nonetheless, the process developed by the authors allows stakeholders in a company to delineate an internal audit procedure that might not unveil all the ethical issues.

Current Challenges in Auditing

In spite of this significant work, the literature on algorithmic auditing presents several challenges. First is the **coordination challenge**. The practices delineated above often focus on specific phases and are insufficient to analyze the entire pipeline of the auditing of an AI system on their own. Thus, auditing complex AI-enabled models require bridging the work of many stakeholders, predicting and addressing a variety of potential harms depending on the context. Therefore, coordination along the phases is key but also complex. In the lifecycle of an AI system, many different stakeholders are accountable for different decisions that impact the overall performance of a system. For instance, the data on which a model is trained significantly affects the performance and accuracy of a model (Gebru et al., 2021). However, the lack of practices to assess whether the data used conforms to the desired deployment contexts and purpose of the system might lead to significant issues, such as reinforcing societal bias by disproportionately affecting vulnerable minorities (Boulamwini & Gebru, 2018).

Secondly is the **implementation challenge**. The translation of regulatory requirements into actionable measures, as well as the delineation of comprehensive auditing frameworks addressing the *entire* lifecycle of a system and guaranteeing independent oversight, is still missing. Measures have to be adopted and defined at an organizational level in order to enable oversight mechanisms and documentation practices that will allow for an understanding of how important decisions were made in the design and development phase of a system and their implications in the deployment phase. Moreover, the implementation challenge is significantly impacted by the current cost of conducting an audit, as expertise in the field is still being built, and the concern of liability exposure caused by the lack of proper measures to detect and mitigate potential issues in the systems creates a significant barrier (Costanza-Chock et al., 2022).

Thus, auditing complex AI-enabled models require bridging the work of many stakeholders, predicting and addressing a variety of potential harms depending on the context.

Another significant issue is the **lack of coordination and collaboration opportunities** between internal and external auditors. Often external auditors face issues around disclosure of the audit results, which leads to a lack of transparency around the potential consequences of deploying a certain system on a large scale. Whereas internal auditing teams face challenges around resources to audit the entire lifecycle of the system developed to unveil all the potential ethical issues (Costanza-Chock et al., 2022; Raji et al., 2020)

Nonetheless, in spite of the significant resources needed to conduct this research, its relevance is tremendous. Developing auditing mechanisms can guarantee lower compliance costs in light of the upcoming regulation, the AI Act, by enabling better mechanisms for conformity assessment. Moreover, auditing mechanisms can also provide greater transparency in terms of an organization's purpose for developing their technology and the efforts are taken to ensure that their systems are safe, fair, and ethically sound in general.

Conclusions

This Brief addresses the fascinating and under-investigated field of algorithmic AI auditing by providing an overview of the practices that have been proposed so far in the field and highlighting some of the challenges that remain in employing these practices. To enable concrete advancement in such an interdisciplinary field of research, however, the need for cooperation between policymakers, AI developers, academia, and other relevant stakeholders is essential.

In light of upcoming regulations on AI governance, AI practitioners are challenged to comply with required rules while lacking standardized practices and documentation tools to conduct technical but also ethical assessments of their AI systems. Hence, interdisciplinary arenas are necessary to enable the development of new methodologies and tools that can help standardize actionable measures to assess the impact of AI systems in a specific domain and mitigate the potential concerning consequences of their deployment in time.

References:

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1

.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Cheong, S. M., Sankaran, K., & Bastani, H. (2022). Artificial intelligence for climate change adaptation. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1459.

Cihon, P. (2019). Standards for AI governance: international standards to enable global coordination in AI research & development. Future of Humanity Institute. University of Oxford.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022, June). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1571-1583).

Dattner, B., Chamorro-Premuzic, T., Buchband, R., & Schettler, L. (2019). The legal and ethical implications of using AI in hiring. Harvard Business Review, 25.

Dattner, B., Chamorro-Premuzic, T., Buchband, R., & Schettler, L. (2019). The legal and ethical implications of using AI in hiring. Harvard Business Review, 25.

Eastwood, N., Stubbings, W. A., Abdallah, M. A. A. E., Durance, I., Paavola, J., Dallimer, M., ... & Orsini, L. (2021). The Time Machine framework: monitoring and prediction of biodiversity loss. Trends in ecology & evolution.

Englund, C., Aksoy, E. E., Alonso-Fernandez, F., Cooney, M. D., Pashami, S., & Åstrand, B. (2021). AI perspectives in Smart Cities and Communities to enable road vehicle automation and smart traffic control. Smart Cities, 4(2), 783-802.

European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. Ethics guidelines for trustworthy AI. PublicationsOffice. https://doi.org/doi/10.2759/177365

European Commission, Directorate-General for Communications Networks, Content and Technology, (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment, Publications Office. https://data.europa.eu/doi/10.2759/791819

European Commission. 2020. White Paper on Artificial Intelligence-A European approach to excellence and trust. Com (2020) 65 Final (2020).

European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM (2021) 206 Final).

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., ... & Yeong, Z. K. (2021). Governing AI safety through independent audits. Nature Machine Intelligence, 3(7), 566-571.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.

IEEE (2018). Algorithmic Bias Considerations. https://standards.ieee.org/ieee/7003/6980/ Accessed July 12, 2022

ISO (2022) Framework for Artificial Intelligence Systems Using Machine Learning https://www.iso.org/standard/74438.html?browse=tc Accessed July 12, 2022

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4).

Knowles, B., & Richards, J. T. (2021, March). The sanction of authority: Promoting public trust in ai. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency(pp. 262-271).

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., ... & Lomas, E. (2021). Towards algorithm auditing: A survey on managing legal, ethical and technological risks of AI, ML and associated algorithms.

Manyika, J., & Sneader, K. (2018). AI, automation, and the future of work: Ten things to solve for.

McKay, C. (2020). Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. Current Issues in Criminal Justice, 32(1), 22-39.

Panch, T., Mattie, H., & Celi, L. A. (2019). The “inconvenient truth” about AI in healthcare. NPJ digital medicine, 2(1), 1-3.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 33-44).

Richards, J., Piorkowski, D., Hind, M., Houde, S., & Mojsilović, A. (2020). A methodology for creating AI FactSheets. arXiv preprint arXiv:2006.13796.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), 1-31.

Yam, J., & Skorburg, J. A. (2021). From human resources to human rights: Impact assessments for hiring algorithms. Ethics and Information Technology, 23(4), 611-623.