

Living with Opaque Technologies: Insights for AI from Digital Simulations

Eugenia Cacciatori^{1,*}, Enzo Fenoglio² and Emre Kazim³

¹Bayes Business School, City, University of London,

²University College London, University of London

³Holistic AI, London

Abstract

This study explores transparency challenges in algorithmic fairness. After reviewing progress in technical and regulatory transparency, we suggest that some level of opacity is inherent to AI systems. Drawing on the relational approach and Polanyi's work on tacit knowledge, we propose studying how society has dealt with other opaque technologies. Using digital simulation modeling as an example, we discuss the similarities and differences between simulations and AI systems in terms of accuracy and transparency. Further research is recommended to advance algorithmic fairness and responsible practices.

Keywords

AI ethics, AI transparency, digital simulations, algorithmic fairness, responsible algorithms.

1. Introduction

There have been significant developments in both the technical properties of a transparent AI system and adequate regulation and legislation. Transparency is increasingly becoming a shorthand to refer to the many aspects of the design, training, and implementation of an AI system needed to ensure that the *whys* and *hows* of AI-based decisions are accessible to humans. This includes issues related to all the elements of an AI system: the data, the system, and the business models [1]. Transparency is also increasingly used in the practice and policy contexts to encompass a variety of technical terms such as explainability, explicability, and interpretability, whose usage and meaning are far from settled [2], but that aim to describe the internal algorithmic logic of an AI system, providing information adapted to the expertise of the stakeholder concerned (e.g., layperson, regulator, or researcher) so that they can perceive it as *transparent*.


In this paper, we propose that useful insights for future research on AI transparency can be drawn from research on the adoption of other *opaque* technologies, in particular digital (or computer) simulations [3, 4].


EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

*Corresponding author.

✉ eugenia.cacciatori@city.ac.uk (E. Cacciatori); e.fenoglio@ucl.ac.uk (E. Fenoglio); emre.kazim@holisticai.com (E. Kazim)

ORCID 0000-0001-6229-7266 (E. Cacciatori); 0000-0003-0224-7237 (E. Fenoglio); 0000-0001-8484-7492 (E. Kazim)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Exploring Transparency Issues in Algorithmic Fairness

2.1. Transparency and Fairness

Fairness, accountability, and transparency are tightly connected in the literature on the ethics of artificial intelligence [5]. Transparency is an important condition for fairness since the perception of fairness depends on the ability to understand the rationale and the process behind a decision. Transparency is also a precondition for accountability. However, fairness is a broader concept than transparency, in particular, because it includes the need for decisions to be free from bias and discrimination but difficult to achieve in practice.

Beyond fairness, transparency is critical for many other functions, such for instance as enabling learning, which in turn improves the design of AI systems [6].

2.2. Current Approaches to Transparency

There are now techniques and methods that can reconstruct, or at least gain some insights, into how an AI system reached its decisions [6]. Several studies have highlighted the potential of post-hoc model-agnostic local explanation methods [1], which focus on explaining individual predictions of any black-box model. Methods, such as **LIME** (Local Interpretable Model-agnostic Explanations) and **SHAP** (Shapley Additive Explanations), have been employed in various domains. Notable examples can be found in high-stakes applications such as healthcare to interpret the decisions made by medical diagnosis models, finance to interpret credit scoring models, or transportation to interpret the decisions made by the self-driving system [7].

Model-agnostic explanations are sensitive to sparsity—most features have no impact, and missing features make explanations difficult. SHAP and LIME differ based on sparsity, and only SHAP excels at identifying important features in sparse regions.

One prominent model-agnostic local explanation is the use of counterfactuals. Counterfactual explanations do not attempt to clarify the internal decision-making process, focusing instead on identifying external factors that could be different to achieve the desired outcome [8]. Because of their nature, counterfactual explanations work around the need to understand the working of the model directly. They are thus a promising way forward that allows the benefit of employing technically opaque but efficient AI systems while also potentially ensuring a level of transparency that is socially acceptable.

From a regulatory point of view, there has been a recognition that transparency is a relational property of AI systems, which emerges from the interaction of particular AI systems in relation to specific issues, specific users, and specific contexts [9]. Thus, the issue of transparency cannot be solved exclusively from technical approaches but requires complex social infrastructures [2] as well.

2.3. The Black Box Issue

While the push towards more transparency is necessary and important, there are good reasons to believe that total transparency is unlikely to be possible. There are epistemological arguments suggesting that complete explainability might not be possible. This viewpoint builds upon the longstanding research tradition, including the influential work of Polanyi [10], which emphasizes

the inherent presence of tacit knowledge within human understanding. Tacit knowledge depends on attention focus, with a conscious part open to scrutiny and a tacit part in the background. Decision-making balances explicit and tacit knowledge. This framework suggests that full transparency and articulation of all knowledge may be unattainable. Recognizing tacit knowledge implies that some aspects cannot be fully expressed or understood. Instead, explanation becomes contextual and relational, shaped by the interplay of tacit and explicit knowledge.

If explanations are relational [9], it is unlikely that we can achieve transparency in every circumstance, and it may not always be essential [11]. As we do not normally require absolute transparency for every technology we use, the widespread usage and established regulatory frameworks have made them an accepted, albeit occasionally controversial, part of our lives. Therefore, overall, there is a general acceptance of treating them as black boxes. A lot of work in regulation and legislation thus aims at creating a similar social infrastructure of trust for AI.

2.4. AI Systems and Mechanistic Tacit Knowledge

Similarly to human knowledge, the knowledge embedded in AI systems is also characterized by tacit elements, so-called *mechanistic tacit knowledge*, which encompasses unobservable and distributed processes in AI systems [2]. The term *Mechanistic* highlights knowledge produced through unknown mechanisms and processes of artificial neural networks (ANNs), not directly observable or manipulable by humans. For example, while engineers program a robot with explicit knowledge (the algorithm) for tasks like riding a bike, the robot's execution relies on inaccessible mechanistic tacit knowledge. The robot lacks explicit knowledge of the algorithm but can still perform the task successfully [12].

The unobservable nature of mechanistic tacit knowledge hinders transparent explanations of AI system behavior, similar to the challenges in explaining human decision-making. This concept is crucial in AI explainability discussions, indicating that certain aspects of AI decision-making may perpetually remain opaque or hard to comprehend [11], akin to human knowledge in modern societies. In addition, even with we had full access to an AI system's internal workings and complete transparency, our comprehension may still be limited because of the challenges of understanding fundamentally different cognitive processes [13]. Thus, integrating mechanistic tacit knowledge and counterfactual explanations provides a promising framework for gaining insights into how AI systems reach decisions.

3. Dealing with Opacity: Directions for Further Research

Further insights into how to address the opacity of AI systems can come from experience with other opaque technologies. Digital simulations are characterized by an *essential epistemic opacity* because “no human can examine and justify every element of the computational processes that produce the output of a computer simulation” [3]. Further, simulations are also opaque because the limited possibilities for experimentation mean that it is often difficult to assess the *truthfulness* of a simulation's result [14], a trait they share with at least some AI systems, such as in healthcare, for which the correctness of decisions is often difficult to assess [15].

Simulations, while opaque, offer valuable problem-solving capabilities. The technical literature today offers extensive normative guidance on how to establish validation and verification procedures, including sensitivity analyses and comparisons with real-world data. Yet, these practices did not emerge fully formed into handbooks—even today, the guidance in handbooks falls well short of accounting for the realities of decision-making through simulations in organizations [4, 16]. Yet, it is these practices that, in the end, determine how opacity is managed and how it impacts how decisions are taken. The literature on simulations suggests that an accounting of the realities of managing AI system opacity in organizations is crucial to make sure that the debate on AI transparency does not remain concerned only with technical issues or a broad regulatory architecture framework, which might have limited or counterproductive effects at worst[17].

The ability to use simulations effectively despite their opacity developed gradually through trial and error processes. The result of these incremental changes, which emerged to accommodate the specific balance between tacit and explicit knowledge afforded by simulation, fundamentally altered the nature of decision-making. For instance, the use of simulations engendered a longstanding debate in science about the nature of the evidence that simulations provide—when can a simulation result be considered evidence for a theory? Over time, a distinctive *mode of knowing* through simulations emerged, with its own rules about which problem can be addressed and what counts as evidence [18]. As AI systems introduce an unavoidable and new (mechanistic) tacit dimension in our decisions processes, we need to investigate the specific organizational practices through which this remaining opacity is managed and how this contributes to reshaping how decisions are reached in organizations.

Finally, as with any other technology, the adoption of simulation is associated with shifts in power between occupations. If a tacit dimension is unavoidable, a debate on transparency within a framework of ethical AI needs to consider how AI systems shift the balance between explicit knowledge, human tacit knowledge, and mechanistic tacit knowledge; and how this changes the nature of decision-making and power balances.

Adapting to the opacity of AI systems won't be straightforward, as simulations indicate. The adoption of AI necessitates new decision-making practices and organizational processes tailored to how it balances tacit and explicit knowledge. These practices may give rise to a new class of professionals. While current transparency tools and regulations will be relevant to shaping these practices, empirical studies are needed to understand the impact of AI on decision-making within individuals and organizations.

4. Conclusion

This paper summarizes transparency approaches and highlights the need to understand the limitations of explaining complex AI systems. We argue that addressing transparency in AI models requires acknowledging the persistence of opacity. Research on digital simulations suggests that this opacity will not fade away with trust infrastructure alone. Adapting to the opacity of AI systems will lead to subtle adjustments in decision-making processes, creating new types of decision-making processes. Research on transparency should engage with these changes to ensure ethical AI deployment in society.

References

- [1] A. Barredo Arrieta, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion* 58 (2020) 82–115. doi:<https://doi.org/10.1016/j.inffus.2019.12.012>.
- [2] E. Fenoglio, E. Kazim, AI Explainability, Interpretability, and Transparency, 2023. (forthcoming).
- [3] P. Humphreys, The philosophical novelty of computer simulation methods, *Synthese* 169 (2009) 615–626. doi:[10.1007/s11229-008-9435-2](https://doi.org/10.1007/s11229-008-9435-2).
- [4] D. E. Bailey, P. M. Leonardi, S. R. Barley, The lure of the virtual, *Organization Science* 23 (2012) 1485–1504. doi:[10.1287/orsc.1110.0703](https://doi.org/10.1287/orsc.1110.0703).
- [5] D. Shin, Y. J. Park, Role of fairness, accountability, and transparency in algorithmic affordance, *Computers in Human Behavior* 98 (2019) 277–284. doi:<https://doi.org/10.1016/j.chb.2019.04.019>.
- [6] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, *WIREs Data Mining Knowl Discov* 11 (2021) e1391. doi:<https://doi.org/10.1002/widm.1391>.
- [7] A. Saranya, R. Subhashini, A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends, *Decision Analytics Journal* 7 (2023) 100230. doi:<https://doi.org/10.1016/j.dajour.2023.100230>.
- [8] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, *Harv. J.L. & Tech.* 31 (2018).
- [9] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* 267 (2019) 1–38. doi:<https://doi.org/10.1016/j.artint.2018.07.007>.
- [10] M. Polanyi, The logic of tacit inference, *Philosophy* 41 (1966) 1–18. doi:[10.1017/S0031819100066110](https://doi.org/10.1017/S0031819100066110).
- [11] B. Brożek, M. Furman, M. Jakubiec, B. Kucharzyk, The black box problem revisited. Real and imaginary challenges for automated legal decision making, *Artificial Intelligence and Law* (2023) 1–14. doi:[10.1007/s10506-023-09356-9](https://doi.org/10.1007/s10506-023-09356-9).
- [12] M. Héder, D. Paski, Autonomous robots and tacit knowledge, *Appraisal* 9 (2012).
- [13] L. J. J. Wittgenstein, *Philosophical Investigations*, New York, US: Wiley-Blackwell, 1953.
- [14] J. Weinkle, J. Roger Pielke, The truthiness about hurricane catastrophe models, *Science, Technology, & Human Values* 42 (2017) 547–576. doi:[10.1177/0162243916671201](https://doi.org/10.1177/0162243916671201).
- [15] S. Lebovitz, N. Levina, H. Lifshitz-Assaf, Is AI Ground Truth Really ‘True’? The Dangers of Training and Evaluating AI Tools Based on Experts’ Know-What, 2021.
- [16] E. Cacciatori, P. Jarzabkowski, R. Bednarek, K. Chalkias, What’s in a Model? Computer Simulations and the Management of Ignorance, *Academy of Management Proceedings* 2019 (2019) 18102. doi:[10.5465/AMBPP.2019.250](https://doi.org/10.5465/AMBPP.2019.250).
- [17] A. Bell, I. Solano-Kamaiko, O. Nov, J. Stoyanovich, It’s Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy, in: *ACM FAccT 2022*, New York, NY, USA, 2022, p. 248–266. doi:[10.1145/3531146.3533090](https://doi.org/10.1145/3531146.3533090).
- [18] M. Morrison, *Reconstructing Reality: Models, Mathematics, and Simulations*, Oxford University Press, 2015. doi:[10.1093/acprof:oso/9780199380275.001.0001](https://doi.org/10.1093/acprof:oso/9780199380275.001.0001).