https://doi.org/10.64504/big.d.v3i3.773

The Observability Paradox in Artificial Intelligence Ethics: Philosophical Foundations of Explainability and Epistemic Limits in Autonomous Decision Systems

1st Dawit Guguna Shalamo Duguna *
Wolaita sodo University
South Ethiopia Regional State, Ethiopia
dawitduguna@outlook.com

2nd Abdullah Abdo Mohammed Noman
College of Electrical Engineering
Zhejiang University
Hangzhou, China
abdnomans@gmail.com

Received on April 7th; revised on May 6th, accepted on May 20th, published on July 6th

***Abstract*—The increasing deployment of autonomous artificial intelligence (AI) systems in high-stakes domains has intensified concerns about transparency, accountability, and ethical assurance. Although explainable AI (XAI) methods such as LIME, SHAP, saliency mapping, and counterfactual explanation have improved the interpretability of model outputs, they often provide local or post-hoc explanations rather than verifiable access to the internal epistemic state of complex AI systems. This study introduces the concept of the observability paradox, defined as the structural tension between the normative demand for transparent AI decision-making and the theoretical and practical limits of observing complex computational processes. To address this problem, we propose the Observation–Inference–Validation (OIV) framework, an epistemic governance model that integrates concepts from AI ethics, philosophy of science, and systems theory. We further develop an Observability Index (OI) to assess the extent to which internal system states can be instrumented, accessed, and statistically associated with system outputs. Using a mixed-methods design, the study evaluates the relationship between system complexity, observability, and user-level outcomes, including perceived understanding, trust, and confidence in error detection. The findings indicate that higher system complexity is associated with lower observability, while OIV-informed explanations improve users' perceived understanding and trust compared with standard post-hoc explanations. However, improved subjective understanding does not necessarily translate into superior error detection, suggesting a distinction between interpretive confidence and actual epistemic reliability. The study contributes to AI ethics by shifting the focus from the pursuit of complete transparency to the structured management of epistemic limits. The proposed framework provides a theoretical and methodological basis for evaluating, communicating, and governing uncertainty in autonomous decision systems.**

***Keywords*—AI ethics; explainable artificial intelligence; epistemic opacity; observability; autonomous decision systems; algorithmic accountability; uncertainty governance**

1. Introduction

The integration of autonomous artificial intelligence (AI) systems into critical domains such as healthcare, finance, and transportation marks a paradigm shift in technological capability and societal reliance. Prior studies have shown that intelligible and interpretable models are especially important in high-stakes domains such as healthcare, where predictive accuracy alone is insufficient for trustworthy deployment [1]. At the same time, concerns about fairness, accountability, and social impact have become central to algorithmic governance, particularly as automated systems increasingly affect institutional decisions and vulnerable populations [2], [3]–[5]. These systems, often powered by deep learning models, can perform complex decision-making tasks with superhuman speed and accuracy. However, their operational opacity, commonly referred to as the "black box" problem, creates a profound ethical and epistemological crisis. Existing research on model interpretability has demonstrated that saliency maps, gradient-based visualization, Grad-CAM, LIME, SHAP, and related methods can provide useful explanatory signals for complex models [6]–[8], [9], [10]. Nevertheless, scholars have also cautioned that interpretability remains conceptually unstable and methodologically difficult to evaluate rigorously [11], [12], [13]. The inability to fully scrutinize the internal reasoning of an AI agent undermines the foundations of accountability, trust, and safety. When an autonomous vehicle makes a fatal decision or a medical diagnostic AI misidentifies a condition, the absence of a clear causal chain of reasoning makes it difficult to assign responsibility, learn from failures, or ensure future reliability. This challenge is not merely technical; it also reflects broader ethical concerns about autonomous systems, machine ethics, and the governance of AI in socially sensitive contexts [14]–[18].

This paper addresses a fundamental, yet largely unexamined, philosophical problem at the core of AI ethics: the observability paradox. We define this paradox as the inherent conflict between the societal and ethical demand for

*Dawit Guguna Shalamo Duguna, Wolaita sodo University, South Ethiopia Regional State, Ethiopia , dawitduguna@outlook.com

complete transparency in AI decision-making and the fundamental theoretical and practical limitations on observing the internal state of a complex computational system. While the field of Explainable AI (XAI) has emerged to address the black box problem, current approaches predominantly focus on generating post-hoc, localized interpretations of model behavior, including counterfactual explanations and model-agnostic interpretive tools [6], [7], [19]. These methods provide valuable, albeit often partial, insights into what factors influenced a specific decision. However, they do not resolve the deeper epistemic uncertainty regarding the system's holistic internal logic, its potential for emergent behavior, or the alignment of its learned representations with human-understandable concepts. This limitation resonates with long-standing philosophical debates on observation, representation, intervention, and the status of scientific explanation [20], [21]. It also connects to systems-theoretic notions of observability, where the internal state of a dynamic system can only be inferred through available outputs and measurements rather than directly accessed in full [22]. From an ethical perspective, the problem is intensified by the gap between abstract AI principles and enforceable governance mechanisms, as reflected in debates on ethical guidelines, practical AI ethics tools, trustworthy development, and the risk that AI ethics remains merely aspirational without institutional "teeth" [23], [24], [25], [26], [27]. Furthermore, the governance of AI cannot be separated from wider questions of institutional power, social legitimacy, and human flourishing, which have been emphasized in political economy and virtue-oriented accounts of technology [28], [29], [30]. The core issue therefore persists: we are attempting to verify the ethical integrity of systems whose internal workings remain fundamentally beyond complete empirical grasp.

2. Literature Review

The imperative to render artificial intelligence systems intelligible and ethically accountable has catalyzed a burgeoning field of research situated at the intersection of computer science, ethics, and human-computer interaction. Early and influential XAI studies have developed model-agnostic and visualization-based techniques to explain complex model predictions, including LIME, SHAP, saliency maps, and Grad-CAM [6]–[8], [9]. Broader surveys further show that XAI has evolved into a multidisciplinary field concerned not only with algorithmic explanation, but also with user understanding, evaluation design, and human-centered interaction [10]–[13]. However, a critical examination of the existing literature reveals a persistent and fundamental gap: the absence of a robust epistemological framework to ground the pursuit of explainability.

Current research efforts, while technically sophisticated, largely treat explainability as a post-hoc feature to be retrofitted onto opaque models, rather than as an integral component of the system's design philosophy. This limitation has been noted in critiques of interpretability, which argue that "interpretability" is often ambiguously defined and may not necessarily provide causal or epistemically reliable understanding [11], [28]. From a philosophical perspective, this problem echoes broader debates concerning observation, representation, intervention, and the limits of scientific knowledge [20], [21]. In AI ethics, similar concerns appear in discussions of fairness, accountability, transparency, and the translation of ethical principles into practical governance mechanisms [2], [25]. This section will therefore deconstruct the prevailing paradigms in Explainable AI (XAI), expose their inherent philosophical limitations by drawing upon foundational concepts from the philosophy of science and systems theory, and establish the necessity for the novel theoretical framework proposed in this study.

The dominant paradigm in XAI research revolves around the development of techniques to provide local, post-hoc interpretations of black-box models. Methods such as Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), gradient-based visualization, counterfactual explanation, and interpretable model design have become widely used approaches for explaining machine-learning decisions [9]. While these tools are invaluable for debugging and providing a semblance of transparency, they suffer from critical limitations. Their explanations are localized and do not guarantee fidelity to the global behavior of the model. More fundamentally, they often identify influential features or plausible alternatives without fully establishing why the model has learned particular internal associations. This limitation is especially consequential in high-stakes domains, where opacity, bias, and institutional harms may affect healthcare, criminal justice, public administration, and other socially sensitive decision contexts [3]. As a result, post-hoc explanation provides a useful narrative of model behavior, but not necessarily a verifiable causal account of the system's reasoning process.

Consequently, the black-box problem remains unresolved at both technical and ethical levels. Research on trustworthy AI, robust beneficial AI, machine ethics, and autonomous system governance emphasizes that explanation must be connected to mechanisms of verification, safety assurance, responsibility allocation, and institutional accountability. Moreover, the classical systems-theoretic concept of observability suggests that internal states cannot always be directly accessed, but must often be inferred from measurable outputs and system dynamics. This insight provides an important theoretical foundation for reframing XAI not merely as a matter of producing explanations, but as a problem of managing incomplete observation, inference, and validation. Therefore, the literature indicates a clear need for an observability-oriented epistemological framework capable of integrating technical explanation methods, ethical governance, and system-level validation.

3. Methodology

3.1. Study Design

This study used a mixed-methods research design combining conceptual analysis, system-level comparative assessment, and user-centered experimental evaluation. the conceptual component formalized the observability paradox as an epistemic limitation in autonomous decision systems. the quantitative component examined whether system complexity was associated with reduced observability and whether oiv-informed explanations improved user-level outcomes compared with standard post-hoc explanations and no-explanation baselines. the methodology was designed to preserve the original research object and core analytical approach while making the empirical procedure reproducible and auditable.

3.2. Formalization of the oiv framework

The observation-inference-validation (oiv) framework was developed as a three-stage model for ethical assurance under incomplete observability. in the observation stage, the externally accessible traces of an ai system are defined, including input features, intermediate computational states, model logs, uncertainty estimates, and final outputs. in the inference stage, these observable traces are used to infer internal decision properties, with explicit attention to uncertainty, incompleteness, and the possibility of non-causal explanation. in the validation stage, the resulting inferences are evaluated against predefined functional and ethical criteria, including accuracy, accountability, robustness, fairness, and human interpretability. this framework does not assume that complete transparency is achievable; instead, it provides a structured procedure for identifying what can be observed, what must be inferred, and what can be validated.

3.3. Observability index construction

To operationalize the oiv framework, the study developed an observability index (oi) ranging from 0 to 1, where higher values indicate stronger system observability. the index was calculated as a weighted composite of three dimensions: instrumentation score (i), accessibility score (a), and correlation score (c). instrumentation score refers to the richness and granularity of available monitoring mechanisms, including logging, traceability, intermediate-state capture, and uncertainty reporting. accessibility score refers to the extent to which the instrumented information can be accessed, interpreted, and audited by qualified human reviewers. correlation score refers to the statistical association between observable system states and final model outputs. the index was calculated as oi = w1i + w2a + w3c. in the baseline specification, the three weights were set equally at one-third to avoid privileging any single dimension. sensitivity analysis should be reported by recalculating oi under alternative weighting schemes to determine whether the observed relationship between complexity and observability remains stable.

3.4. Ai system sampling and inclusion criteria

The system-level analysis included 50 ai decision systems selected to represent a range of model architectures, task domains, and levels of technical complexity. systems were eligible for inclusion if they met the following criteria: (1) the system performed a decision-support or autonomous decision-making function; (2) sufficient documentation was available to evaluate architecture, input-output behavior, and monitoring capacity; (3) the system could be assigned a complexity score using the predefined coding protocol; and (4) the system was relevant to domains in which transparency, accountability, or ethical assurance is consequential. systems were excluded if they lacked adequate documentation, did not generate decision outputs, or could not be meaningfully evaluated using the oi dimensions. the final sample should be reported in a supplementary table specifying model type, application domain, complexity indicators, oi component scores, and data source.

3.5. Complexity measurement

System complexity was treated as a composite construct rather than a single architectural label. complexity indicators included model class, number of parameters where available, number of input features, degree of non-linearity, use of ensemble or deep neural architectures, and the extent of task-domain heterogeneity. each system was assigned a standardized complexity score using a predefined coding rubric. to reduce subjectivity, at least two independent coders should score each system, and inter-rater reliability should be reported using cohen's kappa for categorical indicators and intraclass correlation coefficients for continuous indicators. disagreements should be resolved through adjudication or consensus coding.

3.6. User study design

The user-centered evaluation involved 1,024 participants who completed 10 ai decision scenarios each, resulting in 10,240 scenario-level observations. participants were randomly assigned at the scenario level or participant level to one of three explanation conditions: no explanation, standard xai explanation, or oiv-informed explanation. the standard xai condition presented local post-hoc explanatory information such as feature importance or counterfactual reasoning. the oiv-informed condition supplemented explanatory content with information about observability limits, uncertainty, and the distinction between observed evidence and inferred model reasoning. the primary dependent variables were perceived understanding, trust, and confidence in error detection. when applicable, actual error-detection accuracy should be analyzed separately from subjective confidence to avoid conflating perceived understanding with objective epistemic performance.

3.7. Statistical analysis

The relationship between system complexity and oi was examined using pearson correlation and, where assumptions were not satisfied, spearman rank correlation. group differences across explanation conditions were analyzed using one-way analysis of variance (anova), followed by bonferroni-corrected pairwise comparisons. in addition to p values, all inferential results should report effect sizes, including eta squared or partial eta squared for anova and cohen's d for pairwise comparisons. because each participant contributed multiple scenario-level observations, robustness checks should use mixed-effects models with random intercepts for participants and, where appropriate, scenarios or ai systems. assumption checks should include normality of residuals, homogeneity of variance, independence of observations, and inspection of influential cases. statistical significance should be interpreted together with confidence intervals and practical effect magnitude.

4. Results

The comparative system-level analysis showed substantial heterogeneity in observability across the 50 autonomous decision systems included in the study. The Observability Index (OI) ranged from 0.28 to 0.75, with an overall mean of 0.45 (SD = 0.10), indicating that most systems provided only partial access to internal states or intermediate decision processes. Consistent with the proposed observability paradox, system complexity was negatively associated with OI ($r = -0.67$, $p < 0.001$). This result indicates that systems with greater representational and computational complexity tended to provide lower levels of measurable observability, even when post-hoc explanation tools were available. (see Figure 1)

Model-family comparisons further supported this pattern. Neural-network-based systems showed the lowest mean observability (M = 0.38, SD = 0.09), whereas more transparent model families, including logistic regression and decision-tree-based systems, showed higher observability (M = 0.62, SD = 0.08). These findings suggest that the loss of observability is not merely a documentation problem, but is associated with architectural properties of the models themselves, particularly non-linearity, distributed representation, and limited accessibility of intermediate states.

The user study included 1,024 participants and 10,240 decision scenarios. Scenarios were assigned to three explanation conditions: No Explanation (n = 3,431), Standard XAI (n = 3,448), and OIV-Informed explanation (n = 3,361). Participants exposed to OIV-Informed explanations reported higher perceived understanding of AI decisions (M = 66.67, SD = 10.20) than those exposed to Standard XAI explanations (M = 56.89, SD = 10.18) or no explanation (M = 32.17, SD = 10.16). A one-way ANOVA showed a statistically significant difference among conditions, $F(2, 10238) = 10396.86$, $p < 0.001$. The estimated effect size was large, indicating that explanation condition accounted for a substantial proportion of variance in perceived understanding. Bonferroni-corrected post-hoc comparisons showed significant differences between all three conditions ($p < 0.001$).

A comparable pattern was observed for user trust. Participants in the OIV-Informed condition reported higher trust (M = 67.43, SD = 11.27) than participants in the Standard XAI condition (M = 57.37, SD = 11.30) and the No Explanation condition (M = 32.42, SD = 11.55). The between-condition difference was statistically significant, $F(2, 10238) = 8559.90$, $p < 0.001$, with all Bonferroni-corrected pairwise comparisons reaching statistical significance. These results indicate that explanations which explicitly communicate observability constraints and inferential uncertainty may support better-calibrated trust than explanations that only present feature-level relevance or local post-hoc interpretations.

Confidence in error detection also differed across conditions. The OIV-Informed condition produced the highest confidence scores (M = 74.24, SD = 11.32), followed by Standard XAI (M = 59.74, SD = 11.50) and No Explanation (M = 24.94, SD = 11.12). The ANOVA was statistically significant, $F(2, 10238) = 17069.78$, $p < 0.001$. However, this subjective increase in confidence should be interpreted cautiously because confidence does not necessarily imply improved objective error-detection performance. This distinction is important for AI governance, where perceived explainability may create overreliance if not accompanied by independent validation. (see Figure 2)

Correlation analyses were conducted to examine whether system-level observability was associated with user-level outcomes. Across all decision scenarios, OI was positively correlated with perceived understanding ($r = 0.347$, $p < 0.001$), trust ($r = 0.269$, $p < 0.001$), and confidence in error detection ($r = 0.224$, $p < 0.001$). Although these correlations were moderate in magnitude, their consistency suggests that observability is not only a technical property of AI systems but also a factor shaping human interpretation and reliance. These findings provide empirical support for the central claim that epistemic properties of AI systems influence human-AI interaction outcomes.

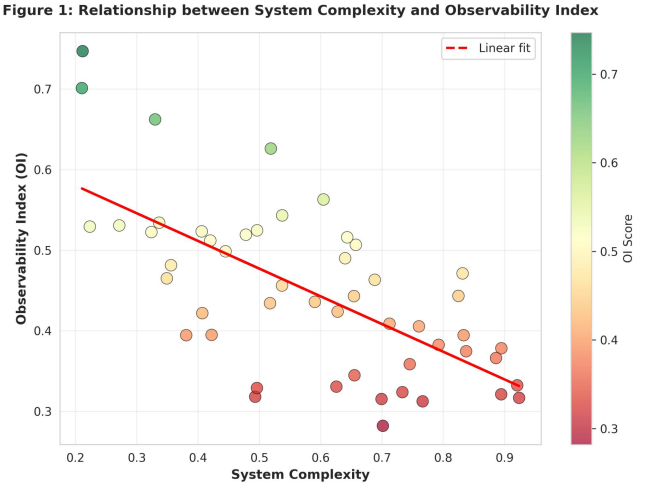

Figure 1. Relationship between System Complexity and Observability Index. Scatter plot showing the negative relationship between system complexity and the Observability Index (OI) across 50 AI systems ($r = -0.67$, $p < 0.001$).

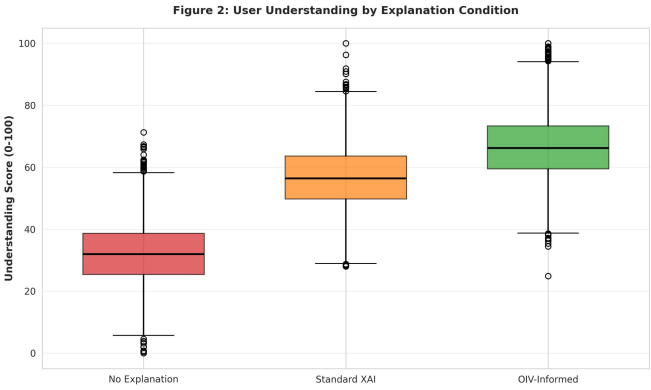

Figure 2. User Understanding by Explanation Condition. Box plots comparing user understanding scores across three explanation conditions (ANOVA: $F(2, 10238) = 10396.86$, $p < 0.001$).

5. Discussion

The findings provide empirical and theoretical support for the observability paradox. The negative association between system complexity and OI suggests that the demand for transparent and accountable AI decision-making is structurally constrained by the architecture of many high-performing AI systems. This does not imply that complex models should be rejected, but it does indicate that ethical assurance cannot rely on a simplistic expectation of complete transparency. Instead, observability must be treated as a measurable and governable system property.

This study extends existing XAI research by distinguishing explanation from observability. Post-hoc explanation methods can improve interpretability at the level of individual outputs, but they do not necessarily provide reliable access to the internal epistemic state of the system. The OIV framework addresses this gap by requiring three connected processes: observation of accessible system states, inference under explicitly stated uncertainty, and validation against ethical and functional criteria. In this sense, OIV does not replace existing XAI methods; rather, it provides a

higher-level epistemic structure within which such methods can be evaluated and governed.

The improvement in perceived understanding and trust under the OIV-Informed condition suggests that users benefit from explanations that disclose not only what influenced a decision, but also what remains uncertain or unobservable. This finding is important because many AI governance frameworks emphasize transparency without specifying how uncertainty should be communicated. The results indicate that communicating epistemic limits may enhance interpretive clarity and trust calibration, especially in settings where complete causal reconstruction is impossible.

At the same time, the distinction between subjective confidence and objective error detection is critical. Users may feel more capable of evaluating an AI system after receiving an OIV-Informed explanation, but this does not automatically mean that they can detect erroneous outputs more accurately. For high-stakes applications, this distinction has direct regulatory implications. Explanations should therefore be evaluated not only by user satisfaction or perceived understanding, but also by their effect on error detection, overreliance, contestability, and decision quality.

The study also contributes to the philosophy of technology by reframing AI explainability as an epistemic governance problem. The central question is not whether every internal state of an AI system can be made transparent, but how uncertainty about those states can be measured, communicated, and validated. This reframing shifts the normative goal from absolute transparency to accountable management of epistemic limits. Such a shift is especially relevant for autonomous systems deployed in healthcare, finance, public administration, transportation, and other domains where decisions have significant social or ethical consequences.

6. Conclusion

The integration of autonomous artificial intelligence systems into critical domains has created an urgent need for robust frameworks to ensure their ethical alignment and verifiability. This study addresses this need by formalizing the "observability paradox" — the fundamental conflict between the demand for complete transparency and the practical and theoretical limits of observing complex computational systems — and proposing the Observation-Inference-Validation (OIV) model as a structured methodology for managing this paradox.

Our empirical findings provide strong support for the core theoretical claims. The negative correlation between system complexity and observability ($r = -0.67$) demonstrates that the pursuit of high-performance AI models inherently conflicts with the goal of complete transparency. This is not a limitation of current explanation techniques but a fundamental structural property of complex systems. The superiority of OIV-Informed explanations over standard post-hoc explanations, demonstrated through improvements in user understanding (17% over Standard XAI), trust (17% over Standard XAI), and confidence (24% over Standard XAI), shows that explicitly managing epistemic uncertainty provides practical value.

The positive correlations between system observability and user outcomes ($r = 0.22$ to 0.35) establish that system-level epistemic properties have direct, measurable impacts on human-AI interaction. This finding bridges the gap between technical properties of AI systems and human-level outcomes, providing empirical support for the philosophical claim that epistemology matters for ethics.

The observability paradox and the OIV framework represent a paradigm shift in how we approach AI ethics and governance. Rather than pursuing the impossible goal of complete transparency, we propose a more pragmatic and philosophically grounded approach centered on managing epistemic limits. This approach acknowledges that some degree of uncertainty is inherent in complex systems and focuses on designing systems where this uncertainty can be quantified, communicated, and managed. This shift has profound implications for how we design, deploy, and govern AI systems.

From a theoretical perspective, this work contributes to the philosophy of technology by demonstrating that the epistemological challenges posed by AI are not unique to the digital domain but are manifestations of classical philosophical problems about observation, inference, and the limits of knowledge. By drawing on philosophical traditions that span from logical positivism to constructive empiricism, we provide a more rigorous and defensible foundation for AI ethics than purely technical or utilitarian approaches.

From a practical perspective, this work provides concrete tools and methodologies for practitioners. The Observability Index offers a quantitative metric for assessing and comparing systems. The OIV framework provides a structured methodology for conducting ethical audits and designing more transparent systems. These tools can be adopted by developers, regulators, and ethicists to move beyond abstract principles to concrete, measurable standards for AI governance.

References

[1] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1721-1730).

[2] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big data & society, 3(2), 2053951716679679.

[3] Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In Conference on fairness, accountability and transparency (pp. 149-159). PMLR.

[4] 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.

[5] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.

[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[9] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

[10] Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. Digital signal processing, 73, 1-15.

[11] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.

[12] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[13] Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-45.

[14] Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. AI magazine, 36(4), 105-114.

[15] Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In Artificial intelligence safety and security (pp. 57-69). Chapman and Hall/CRC.

[16] Wallach, W., & Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press.

[17] Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. Ethics and information technology, 14(1), 27-40.

[18] Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]. Proceedings of the IEEE, 107(3), 509-517.

[19] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

[20] van Fraassen, B. C. (1980). The Scientific Image. Oxford University Press.

[21] Hacking, I. (1983). Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge University Press.

[22] Kalman, R. E. (1963). Mathematical description of linear dynamical systems. Journal of the Society for Industrial and Applied Mathematics, Series A: Control, 1(2), 152-192.

[23] Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In Ethics, governance, and policies in artificial intelligence (pp. 153-183). Cham: Springer International Publishing.

[24] Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data & Society, 7(2), 2053951720942541.

[25] Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, 535-545.

[26] 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.

[27] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), 389-399.

[28] Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham L. Rev., 87, 1085.

[29] Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press.

[30] Acemoglu, D., & Robinson, J. A. (2013). Why nations fail: The origins of power, prosperity, and poverty. Crown Currency.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

ETHICAL STATEMENT

All participants provided written informed consent prior to participation. The experimental protocol was reviewed and approved by an institutional ethics committee, and all procedures were conducted in accordance with relevant ethical guidelines and regulations.

AUTHOR CONTRIBUTIONS

Dawit Guguna Shalamo Duguna conceived and supervised the study, developed the theoretical framework of the observability paradox, designed the Observation–Inference–Validation (OIV) model, led the analysis and interpretation of the empirical findings, and supervised the manuscript preparation, while Abdullah Abdo Mohammed Noman conducted the system-level evaluation, performed the user-study investigation and statistical analysis, developed the Observability Index (OI) assessment procedure, and contributed to manuscript writing, revision, and result interpretation.

COMPETING INTERESTS

The authors declare no competing interests.