Dimensions of Explainability in AI Alignment

Martin Krutský[1], Jiří Němeček[1], Jakub Peleška[1], Paula Gürtler[2]

[1]Dept. of Computer Science, Czech Technical University, Karlovo náměstí 13, 120 00 Prague, Czech Republic
[2]Faculty of Humanities, Charles University, Pátkova 2137/5, 182 00 Prague, Czech Republic

krutsma1@fel.cvut.cz, nemecek.jiri@fel.cvut.cz, jakub.peleska@fel.cvut.cz, guertler@flu.cas.cz

Abstract.

Human-AI alignment is challenging due to limitations in both technical solutions and governance frameworks. Given the infeasibility of properly anticipating all potential misalignment risks, we see explainability as essential for continuous oversight, bridging the gap between AI systems, governance, and human intervention. Recognizing the multi-faceted character of the problem, we argue for a structured framework for evaluating explainability methods, moving beyond narrow technical metrics, to enhance future developments in AI accountability and alignment.

1. Introduction

Ensuring alignment between AI behavior and human values remains a significant challenge in both technical and governance domains. Declarative top-down technical frameworks are widely considered infeasible due to their ambiguity and inability to address real-world complexities [1, 2]—a concern illustrated in literature dating back to Asimov [3], later extended by Bostrom [4]. However, the modern governance approaches encounter analogous difficulties in the implementation and enforcement of the complementary oversight mechanisms and AI regulations, such as the AI Act [5].

Meanwhile, contemporary technical alignment efforts continue to struggle with developing reliable, generalizable solutions for aligning AI systems. While some methods, such as symbolic learning and expert rule-based approaches, offer behavioral guarantees [6], their real-world applicability remains highly limited. Conversely, highly applicable deep learning methods, which rely on data-driven training, exhibit critical failures like reward hacking [7]—highlighting the complementary infeasibility of some safe universal value function optimization. Widespread techniques based on such optimization over user data, such as Reinforcement Learning from Human Feedback (RLHF) and its variants [8, 9, 10], thus remain unreliable. As AI systems operate in increasingly complex environments, unforeseen failure modes are inevitable. Since it is impossible to predefine and test for all risks in advance, continuous human oversight is essential. In this paper, we argue that *explainability* should serve as the foundation for such oversight, enabling ongoing assessment and intervention as new challenges emerge.

Besides ongoing misalignment issues, the development of foundational AI models remains highly centralized, controlled by a few major organizations, limiting external stakeholder influence. This opacity further exacerbates alignment concerns, as stakeholders lack insight into how these models operate. Explainability is, therefore, crucial for *democratizing oversight*, enabling external actors to scrutinize and challenge AI decision-making. Without such transparency, alignment efforts risk becoming monopolized, leading to governance structures that fail to reflect diverse ethical and regulatory perspectives. While initiatives like AI Safety Institutes and emerging directives represent progress, high-level policies alone are insufficient. Effective governance requires complementary bottom-up implementation, particularly in legal contexts where AI decisions must withstand judicial scrutiny. Explainability methods are thus essential to complement top-down regulation with *bottom-up accountability*, ensuring AI behavior remains transparent and justifiable.

Identifying these factors, we argue for a structured, multi-faceted approach to evaluating explainability approaches in the specific context of AI alignment. While there is a broad body of technical explainability research, it prioritizes narrow metrics such as fidelity or attribution accuracy. These techniques fail to capture the broader context of alignment where explainability is not merely a technical concern but a socio-technical bridge between AI systems, governance structures, and human oversight.

2. Assessing XAI methods for Alignment

Despite the abundance of explainability (XAI) methods [11, 12, 13, 14], existing evaluation frameworks focus primarily on technical aspects [15], neglecting the explainability's role in AI alignment and governance. The 12 properties of explanations presented by Nauta et al. [16] are focused on formal requirements (e.g., correctness/faithfulness), practical issues (e.g., compactness), and user experience (e.g., coherence). User experience is (along with trust and performance) also prioritized by Kim et al. [17], Lopes et al. [18], Liao and Varshney [19]. Finally, Islam et al. [20] are proposing high-level dimensions (fidelity, interpretability, robustness, fairness, and complete-

ness), which are, however, evaluated with a very concrete techno-centric metrics and methods.

We argue that effective oversight requires more than technical evaluations and satisfactory user experience—it demands explanations that are understandable to different stakeholders, actionable, and scalable to real-world systems. Crucially, these desiderata should be evaluated jointly to assess their potential trade-offs. Consequently, we propose the following dimensions for XAI methods that not only clarify model behavior but also support human engagement, regulatory auditing, and model corrections.

Intuitiveness to subjects The first dimension is the comprehensibility of explanations to non-expert users. This aligns with regulatory efforts, such as the "right to explanation" [21], which, while not legally binding [22], highlights the importance of intelligibility in AI decision-making. In legal and governance contexts, understandable explanations improve transparency and trust by reducing the need for expert intermediaries. We propose evaluating this criterion through empirical studies, where diverse user groups infer AI decision-making implications based on the explanations. The users' correctness is then the measure of intuitiveness.

Understandability to auditors Beyond lay users, explainability should also support expert evaluation, particularly in regulatory and safety-critical applications. Auditors, policymakers, and AI oversight bodies, such as the AI Office formed under the EU AI Act, require explanations that provide deeper insights into model behavior [23]. Unlike lay users, auditors may possess technical expertise in statistics or machine learning, allowing them to process more complex explanations. Evaluation of this dimension would involve expert assessments of explanation effectiveness in auditing scenarios, such as compliance checks or failure investigations.

Veracity A key requirement of any explanation is that it accurately represents the underlying model's behavior. Often referred to as fidelity or faithfulness [16, 24], this desideratum ensures that explanations are not misleading or oversimplified. For instance, a highly compressed symbolic explanation may be interpretable but fail to reflect the actual decision boundaries of a deep neural network. One possibility is to measure veracity as the expected proportion of inputs for which an explanation remains consistent with the model's true decision function. In local explanations, this would involve testing whether small, explanation-based input modifications yield the expected model changes.

Actionability Explainability methods must enable corrective actions to be useful in AI alignment. If an AI system produces an undesirable outcome, an auditor should be able to infer how to adjust the model. Likewise, an end-user should be advised on how to modify their input to achieve a different, desirable result. Counterfactual and contrastive explanations [25] are a prime example of actionable insights, as they specify minimal changes needed to alter an AI decision. Actionability could be empirically assessed by evaluating whether users or auditors, given an explanation, can successfully adjust the model's behavior in a controlled setting.

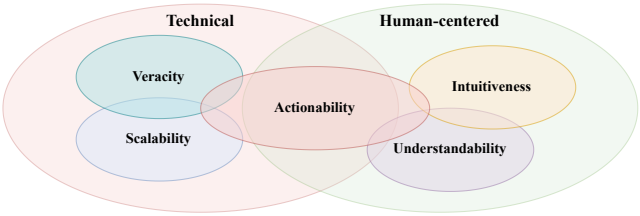

Fig. 1. Relations between the dimensions of explainability.

Scalability Finally, explainability methods must be computationally feasible for modern, large-scale AI systems such as large language models [26]. The increasing complexity of these state-of-the-art models raises concerns about the practical application of interpretability techniques [27]. Scalability could be assessed by measuring the cost of generating explanations across models of varying sizes. A suitable metric would be, e.g., the expected time required to produce explanations under real-world conditions w.r.t. a given level of veracity.

Entanglement of the dimensions The proposed dimensions offer a comprehensive, albeit non-exhaustive, representation of the explainability domain; nonetheless, they are not independent of one another. Figure 1 illustrates the key relationships between the dimensions within both technical and human-centered contexts. We consider intuitiveness and understandability as primary representatives of human-centered perspective, while veracity and scalability emphasize technical aspects. Actionability serves as the connector among all elements—intuitiveness and understandability are intertwined through the empirical measurement of actionability, while scalability and veracity relate to the overall effectiveness of actionability. Additionally, several trade-offs may exist: explanations that are sufficiently simple for lay users may lack the technical depth required by auditors; more intuitive explanations often necessitate simplification, which can compromise veracity; and more accurate explanations frequently demand additional computational resources, thereby impacting scalability.

3. Conclusion

By structuring explainability methods around the proposed dimensions, we aim to foster developments beyond narrow technical metrics towards a broader, multi-faceted explainability research for AI alignment. While the actual assessment of existing XAI methods was out of scope for this work, we laid out these dimensions to highlight the role of explainability in AI alignment—serving not just as a technical tool but as a real bridge between AI systems and humans.

References

[1] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024.

[2] Adam D Thierer. Flexible, pro-innovation governance strategies for artificial intelligence. *R Street Policy Study*, 2023.

[3] I. Asimov. *I, Robot*. Doubleday science fiction. Doubleday, 1950. ISBN 9780385423045.

[4] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., 2014. ISBN 0199678111.

[5] Regulation (EU) 2024/1689. Regulation (EU) 2024/1689 of the European Parliament and of the Council, 2024.

[6] Roberta Calegari, Giovanni Ciatto, and Andrea Omicini. On the integration of symbolic and sub-symbolic techniques for xai: A survey. *Intelligenza Artificiale*, 14(1):7–32, 2020.

[7] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR, 2022.

[8] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26, 2013.

[9] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[10] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[11] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[12] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[13] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

[14] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101, 2024.

[15] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 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, 2021.

[16] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[17] Jenia Kim, Henry Maathuis, and Danielle Sent. Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence*, 7:1456486, 2024.

[18] Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. Xai systems evaluation: a review of human and computer-centred methods. *Applied Sciences*, 12(19):9423, 2022.

[19] Q Vera Liao and Kush R Varshney. Human-centered explainable ai (xai): From algorithms to user experiences, 2021.

[20] Md. Ariful Islam, M. F. Mridha, Md Abrar Jahin, and Nilanjan Dey. A unified framework for evaluating the effectiveness and enhancing the transparency of explainable ai methods in real-world applications, 2024.

[21] Regulation (EU) 2016/679. Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016.

[22] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International data privacy law*, 7 (2):76–99, 2017.

[23] Luca Nannini, Agathe Balayn, and Adam Leon Smith. Explainability in ai policies: A critical review of communications, reports, regulations, and standards in the eu, us, and uk. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1198–1212, 2023.

[24] Miquel Miró-Nicolau, Antoni Jaume-i Capó, and Gabriel Moyà-Alcover. A comprehensive study on fidelity metrics for xai. *Information Processing & Management*, 62(1):103900, 2025.

[25] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[26] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[27] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Acknowledgments

We thank Gustav Šír for his valuable contributions and support in the preparation of this article.

About Authors...

Martin KRUTSKÝ is pursuing PhD in explainable AI and neuro-symbolic learning at the Intelligent Data Analysis Lab at FEE CTU. He has technical expertise in graph neural networks, geometric deep learning, and deep learning explainability. Martin is interested in societal impact of ML and responsible development of AI.

Jiří NEMEČEK's PhD study focuses on AI Explainability and Fairness with guarantees. He utilizes Mixed-Integer Optimization in various applications, from counterfactual explanations to detecting intersectional bias. He is interested in AI's impact on society, including AI Safety/Alignment.

Jakub PELEŠKA is pioneering the field of Relational Deep Learning at the Intelligent Data Analaysis lab at FEE CTU. His PhD study focuses on the machine learning methods applied to relational databases, including graph neural networks.

Paula GÜRTLER is carrying out a PhD in Applied Ethics of AI at Charles University. She investigates the impacts of AI on substance freedoms and how to consider ethics in the design of AI systems.