Explaining Decisions One Conversation at a Time: Opportunities and Risks of LLMs as Explainability Assistants

Filip Cano[a]
Institute of Science and Technology Austria, Klosterneuburg, Austria
filip.cano@ist.ac.at

Keywords: Explainable AI, Large Language Models, Trust in AI

Abstract: Modern AI systems increasingly rely on opaque, highly complex models whose inner workings remain inaccessible even to experts. This opacity creates challenges for trust, accountability, and compliance with emerging regulatory expectations such as the "right to an explanation". While traditional explainability methods—feature attributions, counterfactuals, surrogate models—and interpretable model classes provide valuable insights for engineers, they often fall short of delivering the contextual, conversational explanations that real users expect. Large Language Models (LLMs) offer a promising new avenue for explanation due to their ability to engage interactively, adapt to user needs, and translate technical outputs into more accessible reasoning. However, their tendencies toward hallucination, conflict avoidance, and oversimplification introduce serious risks when used as explanatory agents. This paper analyzes these opportunities and limitations, examines verification strategies for ensuring explanation fidelity, and situates LLM-generated explanations within broader concerns about public trust. The paper concludes by outlining best practices and future research directions for building robust, verifiable, and human-aligned explanation systems.

1 Introduction

Modern AI systems, particularly large neural models, have reached a level of complexity that vastly exceeds human ability to fully grasp their internal mechanisms. These systems now underpin decisions in domains such as healthcare, finance, criminal justice, transportation, and public administration (Lipton, 2018; Esteva et al., 2017; Berk et al., 2019; Khandani et al., 2010). Their deployment in such high-stakes settings raises profound concerns: if even experts cannot precisely articulate why a model behaves as it does, how can individuals, institutions, or regulators trust its outputs?

As societies increasingly rely on opaque algorithms to make consequential judgments, the question of trust becomes central. Without the ability to scrutinize or meaningfully interrogate an AI system, users must take its conclusions on faith. This "trust without understanding" is at odds with the expectations we place on critical infrastructures and decision makers. As a result, pressure is mounting for AI systems to provide reasons for their predictions (European Union, 2024).

Explainable AI. Explainability techniques have emerged to illuminate the behavior of black-box models and help engineers debug, validate, and improve them. Tools such as feature attributions (Zhou et al., 2022), saliency maps (Alqaraawi et al., 2020), and counterfactual explanations (Verma et al., 2024) allow practitioners to diagnose model failures, verify adherence to design intent (Cano Córdoba et al., 2023), and ensure robustness (Chander et al., 2025). From this technical standpoint, explainability is a fundamental component of responsible engineering.

Right to an explanation. But explainability also has a normative and legal dimension. In domains governed by fairness, accountability, and transparency requirements, individuals affected by algorithmic decisions increasingly claim a "right to an explanation." (Edwards and Veale, 2018; Kaminski, 2021). This expectation is reflected in regulatory frameworks such as the GDPR and recent AI governance proposals (European Union, 2024; HLEG AI, 2019; Voigt and Von dem Bussche, 2024). Explanations, in this context, serve to justify actions, empower contestation, and ensure procedural legitimacy. Thus, explainability is not merely an engineering convenience, it is a societal demand (Selbst and Powles, 2018; Wachter

[a] https://orcid.org/0000-0002-0783-904X

et al., 2017a; Barocas and Selbst, 2016).

Interpretable AI. Interpretability is a different path towards providing explanations, by designing models that are inherently understandable to humans, such as decision trees, rule lists, or linear models (Graziani et al., 2023). Unlike explainability, which typically involves post-hoc analysis of opaque systems, interpretability emphasizes transparency by construction. In principle, interpretable models allow users to directly inspect the logic behind decisions and therefore enable trust grounded in clarity rather than reconstruction. Both approaches contribute to trust, but they differ in their epistemic commitments: interpretability aims for simplicity and legibility, while explainability aims for intelligibility even in intricate systems. The interplay between the two is central to debates about how to build AI systems that can be relied upon.

What is an explanation? For both explainable and interpretable AI, a growing critique is that the right to an explanation cannot be satisfied by a single numerical attribution method or by presenting users with a simplified surrogate model. Shapley values, feature importances, or decision-tree approximations may assist experts, but they fall short of what ordinary individuals typically seek: a meaningful, contextual, and interactive account of why a decision was made (Poursabzi-Sangdeh et al., 2021; Kleinberg et al., 2018). Satisfying explanations are seldom static artifacts, they are a dialog. Humans understand through conversation, asking follow-up questions, seeking clarifications, and exploring alternative scenarios (Miller, 2019; Lombrozo, 2006).

Counterfactual explanations capture part of this dialogic nature by showing users how specific changes would alter the outcome. Yet even counterfactuals are limited: they are predefined, fixed in form, and often lack awareness of the user's concerns or background knowledge. Large Language Models (LLMs), by contrast, can engage in open-ended discourse, adapt to the user's level of understanding, and integrate technical explanations with conceptual reasoning (Brown et al., 2020; Touvron et al., 2023). Their conversational flexibility positions them as promising candidates for generating explanations that feel genuinely interactive and human-centered.

LLMs as explainability agents. In this paper, we argue that LLMs offer a great potential to satisfy this right of an explanation that AI users have. We review the emerging approaches that use LLMs as mediators, translators, or meta-explainers for other AI systems, identifying current trends and common techniques. At the same time, we scrutinize the risks of relying on LLMs as explanatory agents. Their tendency to hallucinate, avoid confrontation, or oversimplify complex ideas poses serious challenges, especially in sensitive domains. We discuss issues of reliability, epistemic authority, and alignment, and explore how these systems might be evaluated or constrained. Building on this analysis, we outline future research pathways for designing LLM-driven explanation systems that are trustworthy, transparent, and aligned with human expectations.

Outline of the paper. This position paper is structured as follows. In Section 2 we introduce the main concepts of explainable and interpretable AI. In Section 3 we define the two primary objectives of explanations. The next three sections delve into the potential risks of LLM-based explanations and how to mitigate them: in Section 4 we analyse the potential effect of known LLM malfunctioning sources, like hallucination; in Section 5 we explore the possibilities for explanations to be verified against a real model; and in Section 6 we discuss trust in AI from a more sociological perspective. Section 7 reviews the current uses of LLMs as explanation assistants in the literature. We wrap-up the paper with some concluding remarks on best practices (Section 8) and avenues for future work (Section 9).

2 Background

Before the rise of deep learning, the predominant machine learning models were intrinsically interpretable (Breiman, 2001; Hastie and Tibshirani, 1986). Linear regression, logistic regression, naïve Bayes classifiers, and small decision trees allowed practitioners to examine coefficients, probabilistic parameters, or rule structures directly. These models were favored both because of their computational efficiency and because their internal logic was transparent, making it straightforward to trace how input features influenced predictions. Early work in knowledge-based systems similarly stressed explicit symbolic representations, such as expert systems with human-readable rules, so that reasoning steps could be inspected and critiqued (Clancey, 1983; Akerkar and Sajja, 2009).

The shift toward more complex models, like support vector machines with kernel functions, ensemble methods, and eventually deep neural networksm, marked a rupture in this tradition (Breiman, 2001). Although these models achieved unprecedented accu-

racy, they sacrificed transparency. The "black-box" nature of their internal representations motivated the emergence of explainability and interpretability research as attempts to restore some form of intelligibility without relinquishing performance (Ribeiro et al., 2016; Lundberg and Lee, 2017).

Interpretable vs. Explainable AI. A central conceptual distinction in the literature is that interpretability refers to models whose structure is inherently understandable to humans, whereas explainability refers to techniques applied after the fact to extract insights from opaque models. Interpretable models aim for transparency by design: their parameters or rules can be inspected and comprehended without auxiliary tools. In contrast, explainability methods typically operate by probing, approximating, or analyzing the outputs of complex models to generate post-hoc explanations that approximate the reasoning process.

2.1 Explainability Methods

Global vs. Local Explainability. Explainability methods can be broadly categorized into global and local approaches (Dwivedi et al., 2023). Global explainability aims to describe the overall behavior of a model: its decision boundaries, feature dependencies, or structural patterns. Such methods provide a high-level understanding of how a model behaves on average across the input space, but often struggle to capture fine-grained details for specific instances, especially in high-dimensional or non-linear systems (Rudin, 2019). Local explainability, by contrast, focuses on explaining individual predictions. Techniques such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and gradient-based attribution methods (Ancona et al., 2019) provide instance-level insights by highlighting which features influenced a particular decision. Local explanations are particularly useful in high-stakes scenarios where individuals demand justification for specific outcomes. However, local and global perspectives may diverge: highly accurate local explanations might not generalize globally, and global summaries may obscure crucial local behaviors.

Surrogate models. Surrogate models are among the most widely adopted explainability tools. They operate by training an interpretable model, typically a shallow decision tree or a linear model, to approximate the behavior of a complex black-box model either globally or locally. The surrogate is easier to understand, and its structure can reveal approximate decision boundaries or feature dependencies present in the underlying model. However, surrogate models raise important epistemic concerns, as they may misrepresent the true logic of the underlying system. Their fidelity must be evaluated carefully, and even high-fidelity surrogates may fail to capture subtle interactions or rare behaviors.

Counterfactual explanations Counterfactual explanations describe how a prediction would change if certain input features were altered (Guidotti, 2024; Verma et al., 2024). They answer questions of the form: "What is the smallest change needed for a different outcome?" This framing directly resonates with human reasoning, as people often explain phenomena by contrasting them with close alternatives. Counterfactuals are especially appealing in domains where actionable guidance is valuable. The literature on counterfactual explanations has grown rapidly, encompassing optimization-based techniques, causal-model-based approaches, and methods designed to ensure feasibility or fairness (Wachter et al., 2017b). Despite their intuitive nature, counterfactuals can suffer from issues such as invalid or unrealistic suggestions, multiple equally valid alternatives, and blind spots regarding the model's internal structure (Prado-Romero et al., 2024).

2.2 Interpretability Methods

Decision trees. Decision trees represent one of the clearest examples of interpretable models. Their hierarchical structure mirrors human decision-making: each internal node corresponds to a feature test, and each leaf corresponds to a prediction. This structure allows users to trace the logic of a prediction through a sequence of easily understood rules. Variants such as CART (Breiman et al., 2017), C4.5 (Hssina et al., 2014), and Bayesian decision trees (Linero, 2017) have been extensively studied and remain staples of interpretable machine learning. Despite their transparency, decision trees face limitations of interpretability when high-dimensional or irregular datasets require large trees to achieve accurate performance. As a result, research has focused on learning compact, constrained, or rule-based variants that preserve transparency while maintaining competitive accuracy.

Generalized additive models. Generalized Additive Models (GAMs) offer a powerful balance between flexibility and interpretability (Hastie and Tibshirani, 1986). By expressing predictions as a sum

of feature-wise functions, GAMs allow users to understand the contribution of each variable independently while still capturing non-linear relationships. Modern variants, such as Explainable Boosting Machines (Lou et al., 2013) and Neural Additive Models (Agarwal et al., 2021), extend this framework using boosted or neural components while preserving the additive structure that makes GAMs transparent. As a result, GAMs have become one of the most successful interpretable model families, often achieving accuracy close to black-box models in structured-data domains while remaining inherently understandable (Nori et al., 2021; Chang et al., 2021).

Rule-based and sparse logical models. Rule-based models constitute another major class of interpretable architectures. Unlike decision trees, which organize tests hierarchically, rule lists and decision sets represent a model as a compact collection of human-readable if–then statements. Approaches such as Bayesian Rule Lists (Yang et al., 2017), CORELS (Angelino et al., 2018), and decision sets (Lakkaraju et al., 2016) aim to produce sparse rule structures that remain faithful to the data while maintaining comprehensibility. These symbolic representations are appealing because they resemble the structure of human reasoning and allow straightforward inspection of the logic underlying predictions.

2.3 Natural Language Processing for Explainable AI

The intersection of explainability and natural language generation has a long history (Reiter and Dale, 1997), up to the current interest in large language models. Early work in expert systems included rule-based natural language explanations that verbalized inference chains (Swartout, 1983; Clancey, 1983). As data-driven models grew in prominence, researchers developed template-based or retrieval-based approaches to generate justifications or descriptions, particularly in recommendation systems and knowledge-based reasoning (Tintarev and Masthoff, 2010).

With the advent of neural NLP, sequence-to-sequence models enabled free-form generation of explanations conditioned on model inputs or internal representations (Sutskever et al., 2014; Bahdanau et al., 2015). This led to lines of research on rationalizing predictions, generating faithful explanations tied to attention weights, and training models to output both predictions and textual justifications. Although these early methods were limited in fluency and adaptability, they laid the conceptual foundation for using language as an explanatory medium (Lei et al., 2016). The rise of large pretrained language models dramatically extended this potential, enabling interactive, dialogic explanations not possible with earlier approaches (Devlin et al., 2019).

3 The Two Goals of Explanations

Explanations serve two primary functions within the broader ecosystem of AI systems: they support engineering goals and they fulfill human-centered requirements. These two perspectives, although related, lead to different expectations, different standards of adequacy, and sometimes conflicting desiderata. Understanding this duality is essential for assessing what kinds of explanations are appropriate in different contexts and what role emerging methods, including LLM-based interactive explanations, can play.

The engineer's perspective. From the perspective of the AI practitioner, explanations function primarily as tools for analysis, validation, and debugging. Interpretable models provide immediate insights into how inputs influence outputs, enabling practitioners to diagnose errors, detect spurious correlations, and assess robustness. Post-hoc explainability methods complement this by offering ways to interrogate complex black-box systems that cannot be directly inspected. Feature attributions, counterfactuals, and surrogate models help engineers uncover failure modes, verify alignment with domain knowledge, and ensure that the system behaves consistently across different regions of the input space. In this sense, explanations are integral to the engineering lifecycle of AI systems: they help developers maintain, improve, and eventually trust the systems they build.

The subject's perspective. For individuals affected by algorithmic decisions, explanations play a fundamentally different role. Here, the goal is not debugging but justification, accountability, and procedural fairness. Subjects often expect explanations that are understandable, context-sensitive, and meaningful in terms of their lived experience. While interpretable models may offer transparency, they are not necessarily accessible: even simple rule structures or linear models can be challenging for laypersons without technical expertise. Moreover, high-transparency models can sometimes be exploited or "gamed", raising concerns about deliberate manipulation in strategic environments.

Post-hoc explainability methods promise to bridge this gap, but they also present risks. Some explanations may be misleading, incomplete, or overly technical; others may fail to capture the actual causal factors behind a decision. For this reason, effective explanations for subjects must do more than expose model internals: they must engage with the user's goals, comprehension level, and rights. The "right to an explanation" is therefore best understood not merely as access to technical artifacts, but as access to communicative, person-centered reasoning.

A loan application example. Consider a loan applicant who is denied credit by an automated system. For the engineer, an explanation such as a feature-attribution plot or a counterfactual query may reveal that the model has learned an unintended dependence on geographic information or an unstable interaction between income and employment history. This insight guides debugging and model improvement. For the subject, however, the same explanation may be unhelpful or unintelligible. What they need is a clear and actionable account clarifying why the decision was made and what changes might alter the outcome.

4 The Risk of Explanations that Just Sound Right

Hallucinations. Large Language Models have demonstrated remarkable fluency and flexibility in generating natural-language explanations, but these strengths come with structural weaknesses. Chief among them is their propensity to hallucinate, i.e., to produce statements that are syntactically coherent yet factually incorrect or entirely fabricated (Ji et al., 2023). When LLMs are framed as explainers of other AI systems, this risk becomes particularly acute. An explanation that is eloquent but false can be more harmful than no explanation at all, because users may mistake confident presentation for correctness.

Human interaction. Furthermore, LLMs often exhibit conflict avoidance or excessive deference, tailoring their responses to maintain conversational harmony rather than faithfully representing the underlying reality (Perez et al., 2023). This tendency to social aligment can distort explanations by suppressing ambiguity or disagreement that would otherwise be crucial for understanding model limitations.

Oversimplifcation. Additionally, LLMs are known to oversimplify complex models or decisions when prompted for explanations. While simplification can enhance accessibility, it also invites epistemic danger: intricate dependencies, nonlinear interactions, and edge-case behaviors may be glossed over in favor of a narrative that flows well and seems correct at first sight. These tendencies raise fundamental concerns about whether LLM-generated explanations can be trusted without additional safeguards. If explanations are allowed to drift from the true computational mechanisms of the system being explained, they risk becoming a form of rationalization rather than genuine insigh, a phenomenon already documented in earlier explanation-generation models (Wiegreffe and Pinter, 2019; Lombrozo, 2006). Thus, deploying LLMs as explanatory agents requires careful attention to failure modes, robustness, and the broader implications of delegating epistemic authority to inherently fallible generative systems.

5 Verification of Explanations

Given these risks, the verification of explanations becomes an essential component in any system that relies on LLMs to interpret or translate the reasoning of another model. Ideally, explanations should be directly anchored to verifiable properties of the underlying model, be it through querying the model, accessing its internal representations, or running controlled counterfactual experiments. In settings where the model is fully accessible, verification can be systematic: for example, one can test whether a purported feature dependency aligns with actual model behavior, or whether a claimed counterfactual condition indeed changes the model's prediction. Such checks help ensure that explanations are faithful rather than merely plausible.

However, verification becomes substantially more difficult when the AI system being explained is itself opaque or proprietary, or when explanations must be generated in real time for interactive use. In such cases, approximate or probabilistic verification methods (Yeh et al., 2019; Slack et al., 2020) may be needed, such as evaluating explanation consistency across multiple perturbations or comparing LLM-generated explanations with independently derived attribution signals. There is also a need for metrics that can measure both the informativeness of explanations as well as their fidelity to the underlying model. As LLM-based explanation systems grow in prominence, the development of robust verification pipelines will be crucial to prevent the emergence of ungrounded or misleading insights.

6 The Evolution of Trust in AI

Momentum in public trust. Public trust in a technology does not evolve linearly: it is shaped by cultural momentum, narratives, and high-profile successes or failures. We live in an era characterized simultaneously by unprecedented access to information and by widespread misinformation. In this context, the public perception of LLMs is paradoxical: a mixture of fascination at their capabilities and apprehension about their potential impact. This fragile combination of awe and unease forms the background against which AI systems must earn public trust.

History shows that reputational damage in AI tends to be enduring. The case of COMPAS (Kirchner et al., 2016) is emblematic: although the system was only one of many risk-assessment tools, the controversy surrounding it became a touchstone for discussions of algorithmic bias. As a consequence, public willingness to trust algorithmic decision-makers in judicial settings has been profoundly weakened (Dressel and Farid, 2018; Selbst, 2017). These "trust stains" are difficult to remove once they become embedded in collective memory.

Responsible deployment. This poses a critical challenge for the future of AI, especially as LLM-based explanation systems begin to mediate how users understand algorithmic decisions. If such systems deliver misleading or fabricated explanations, they risk generating the same kind of irrevocable reputational damage. In high-stakes settings, a single widely publicized failure could undermine trust not only in a specific system but in the broader idea that AI can explain itself at all. Therefore, it is imperative to design explanatory frameworks that are robust, verifiable, and transparent about their limitations. Protecting public trust is not simply a matter of system performance; it is a matter of ensuring that the narratives surrounding AI remain grounded in truth and responsibility.

7 The Current Use of LLMs for Explanations

A survey on the use of LLMs as explainers can be found in (Bilal et al., 2025). Current applications of LLMs as explanatory tools largely fall into two categories: assistive interpretation of existing AI models and direct generation of natural-language justifications. In the first category, LLMs are used to summarize model behaviors, interpret saliency or attribution maps, translate counterfactuals into accessible language, or act as conversational interfaces layered on top of technical explainability methods (Ding et al., 2025; Spitzer et al., 2024). Early empirical studies suggest that such systems can increase user satisfaction and perceived clarity, even when the underlying explanations remain approximate (Lubos et al., 2024). In the second category, LLMs are used to directly generate rationales for their own predictions or for the predictions of other models (Wei et al., 2022). While these justifications often appear coherent and contextually relevant, research has shown that they may not faithfully reflect the model's reasoning (Turpin et al., 2023). As a result, LLMs are currently used with caution: they hold promise as mediators of explanations, but they are not yet reliable enough to serve as standalone explanatory agents in high-stakes settings.

8 Best Practices

We propose five guiding principles for the use of LLMs as explanation systems.

1. *Ground explanations in verifiable model behaviour.* The LLM should be guided by signals from the underlying model, preventing it from generating free-form narratives without being anchored in reality.
2. *Clear communication of uncertainties and limitations.* Users should not be led to believe that LLM-generated explanations are infallible, and should be actively reminded of these limitations.
3. *Adapt to your audience.* What is suitable for engineers may overwhelm or frustrate lay users, and vice versa.
4. *Leverage multi-method corroboration.* Whenever possible, combine LLM explanations with traditional explainability tools such as feature attributions, counterfactuals, or rule-based summaries.
5. *Transparency and auditability.* The explanation pipeline should allow failures or inconsistencies to be identified and corrected.

9 The Road Ahead

Future research should focus on building LLM-based explanation systems that are both faithful and adaptively conversational. A key direction is integrating LLMs with the internal representations of the models they explain, enabling hybrid systems where fidelity can be dynamically verified. Another priority is developing evaluation metrics that assess not only linguistic quality but also epistemic correctness and user

impact. Finally, researchers must examine societal and ethical implications, particularly how to prevent erosion of public trust.

Acknowledgements

This work has been supported by the European Research Council under Grant No.: ERC-2020-AdG 101020093. LLM–based tools have been used as writing assistance to help improve presentation.

REFERENCES

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. (2021). Neural additive models: Interpretable machine learning with neural nets. In *Proceedings of NeurIPS 2021*.

Akerkar, R. and Sajja, P. (2009). *Knowledge-based systems*. Jones & Bartlett Publishers.

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the IUI 2020*.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2019). *Gradient-Based Attribution Methods*, pages 169–191. Springer.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18(234):1–78.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671.

Berk, R., Berk, D., and Drougas, D. (2019). *Machine learning risk assessments in criminal justice settings*. Springer.

Bilal, A., Ebert, D., and Lin, B. (2025). LLMs for explainable AI: A comprehensive survey. *arXiv preprint arXiv:2504.00125*.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (2017). *Classification and regression trees*. Chapman and Hall/CRC.

Brown, T., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language models are few-shot learners. In *Proceedings of NeurIPS 2020*.

Cano Córdoba, F., Judson, S., Antonopoulos, T., Bjørner, K., Shoemaker, N., Shapiro, S. J., Piskac, R., and Könighofer, B. (2023). Analyzing intentional behavior in autonomous agents under uncertainty. In *Proceedings of IJCAI 2023*.

Chander, B., John, C., Warrier, L., and Gopalakrishnan, K. (2025). Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness. *ACM Computing Surveys*, 57(6):1–49.

Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A., and Caruana, R. (2021). How interpretable and trustworthy are gams? In *Proceedings of KDD 2021*.

Clancey, W. J. (1983). The epistemology of a rule-based expert system—a framework for explanation. *Artificial Intelligence*, 20(3):215–251.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*.

Ding, S., Vasa, S., and Ramadwar, A. (2025). Explanation-driven counterfactual testing for faithfulness in vision-language model explanations. In *NeurIPS 2025 Workshop on Regulatable ML*.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., et al. (2023). Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33.

Edwards, L. and Veale, M. (2018). Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? In *Proceedings of IEEE Security & Privacy (SP) 2018*.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

European Union (2024). The EU artificial intelligence act.

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J. P., et al. (2023). A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, 56(4):3473–3504.

Guidotti, R. (2024). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 1(3):297–310.

HLEG AI (2019). Ethics guidelines for trustworthy AI. *European Commission*.

Hssina, B., Merbouha, A., Ezzikouri, H., and Erritali, M. (2014). A comparative study of decision tree id3 and c4. 5. *International Journal of Advanced Computer Science and Applications*, 4(2):13–19.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Kaminski, M. E. (2021). The right to explanation, explained. In *Research handbook on information law and governance*, pages 278–299. Edward Elgar Publishing.

Khandani, A. E., Kim, A. J., and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787.

Kirchner, L., Mattu, S., Larson, J., and Angwin, J. (2016). Machine bias. *ProPublica*.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of KDD 2016*.

Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. In *Proceedings of EMNLP 2016*.

Linero, A. R. (2017). A review of tree-based bayesian methods. *Communications for Statistical Applications and Methods*, 24(6).

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

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of KDD 2013*.

Lubos, S., Tran, T. N. T., Felfernig, A., Polat Erdeniz, S., and Le, V.-M. (2024). LLM-generated explanations for recommender systems. In *Proceedings of UMAP*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of NeurIPS 2017*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Nori, H., Caruana, R., Bu, Z., Shen, J. H., and Kulkarni, J. (2021). Accuracy, interpretability, and differential privacy via explainable boosting. In *Proceedings of ICML 2021*.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2023). Discovering language model behaviors with model-written evaluations. In *Proceedings of ACL 2023*.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of CHI 2021*. ACM.

Prado-Romero, M. A., Prenkaj, B., Stilo, G., and Giannotti, F. (2024). A survey on graph counterfactual explanations: definitions, methods, evaluation, and research challenges. *ACM Computing Surveys*, 56(7):1–37.

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you? explaining the predictions of any classifier. In *Proceedings of KDD 2016*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Selbst, A. and Powles, J. (2018). Meaningful information and the right to explanation. In *Proceedings of FAccT 2018*.

Selbst, A. D. (2017). Disparate impact in big data policing. *Georgia Law Review*, 52:109.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of AIES 2020*.

Spitzer, P., Celis, S., Martin, D., Kühl, N., and Satzger, G. (2024). Looking through the deep glasses: How large language models enhance explainability of deep learning models. In *Proceedings of Mensch und Computer 2024*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of NeurIPS 2014*.

Swartout, W. R. (1983). XPLAIN: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21(3):285–325.

Tintarev, N. and Masthoff, J. (2010). Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, pages 479–510. Springer.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Turpin, M., Michael, J., Perez, E., and Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proceedings of NeurIPS 2023*.

Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., and Shah, C. (2024). Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12):1–42.

Voigt, P. and Von dem Bussche, A. (2024). *The EU general data protection regulation (GDPR): A practical guide*. Springer.

Wachter, S., Mittelstadt, B., and Floridi, L. (2017a). 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.

Wachter, S., Mittelstadt, B., and Russell, C. (2017b). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS 2022*.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of EMNLP 2019*.

Yang, H., Rudin, C., and Seltzer, M. (2017). Scalable bayesian rule lists. In *Proceedings of ICML 2017*.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. (2019). On the (in) fidelity and sensitivity of explanations. In *Proceedings of NeurIPS 2019*.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. (2022). Do feature attribution methods correctly attribute features? In *Proceedings of AAAI 2022*.