

Explainable and Ethical Artificial Intelligence Using Language Models

Tanmay Milind Bhole, Dr. Faizur Rashid

School of Computational Sciences, JSPM University Pune, Maharashtra, India

Professor, School of Computational Sciences, JSPM University Pune, Maharashtra, India

Abstract - Increasingly as machine learning systems are used in important areas such as healthcare, finance, criminal justice and government, developers will be required to develop systems that can be examined and queried by those not involved in developing them (whereas before developers typically only dealt with data professionals). This paper pulls together two separate streams of research. The first stream provides an overview of strategies for ensuring human involvement and control over AI systems based on results from surveys of over 160,000 individuals and recommendations made in publications from a journal called Technology in Society [1]. The second stream provides empirical evidence for GPT-based self-correcting natural language explanation pipelines applied to criminal risk assessment data (i.e., an example of applying machine learning to obtain factual verification for GPT-4 at the level of 0.996) [2]. In addition, this paper provides a review of existing post-hoc explainability techniques (e.g., LIME [12], SHAP [11]), reviews the rapidly growing body of literature focused on natural language generated by LLMs [3]–[10] and discusses the problem of fluency bias associated with LLM-generated text. Collectively, these streams offer a framework for successfully executing the principles of explainable AI through architectural quality management processes and socio-technical governance frameworks.

Key Words: Explainable AI, XAI, large language models, GPT, LIME, SHAP, natural language explanation, AI governance, human oversight, responsible AI, confabulation, trustworthy AI.

1. INTRODUCTION

We've been able to develop some amazing systems due to machine learning. However, many of these systems are black boxes that are not accessible or understandable by their users. For example, a deep neural network that has been trained on millions of examples can produce results that are equal to or better than those of a human being. However, the person using the deep neural network doesn't have an understanding of why it produces those results [1][2]. The computation associated with

programming a deep learning network consists of hundreds of millions of nonlinear interactions between parameters, and the results the system produces do not have a direct basis for establishing a cause-and-effect relationship that a clinician or judge could use to explain how they arrived at a decision that affects a person's life [3]. The need to make the output of automated systems understandable has been the primary driver for the development of Explainable Artificial Intelligence (XAI). In the past ten years, researchers have created a broad array of techniques to make the output of machine-learning systems understandable to the widest audience possible without losing any of the predictive power of the models [4][5]. However, making the models comprehensible to a data scientist does not equate to making the models comprehensible to a clinician, a judge or a layperson whose lives may depend on an automated decision being rendered [6]. The divide between technical explain ability and real-world accessibility represents one of the most significant, recurring unresolved challenges in this field. Large language models (LLMs) like GPT-4 could bridge the gap between experts and the general public by translating technical results into simple language. Therefore, non-expert users can read, ask questions about, and challenge AI-generated decisions (2, 9). However, there are trade-offs associated with this benefit. LLMs cannot truly access the internal workings of a classifier; rather, the explanations generated by LLMs are constructed from post hoc patterns in training data. Misleadingly confident and polished narratives can mislead users and lead to incorrect decisions; hence, users may be misled by a story than by a short, effectively numeric response (10, 13). This article investigates these two related issues. It integrates a governance perspective of keeping AI under human control (1) and the results of a study evaluating the ability of GPT-generated self-correcting pipelines to generate English language explanations (2) into the context of the broader body of the Explainable AI literature (3-25). These two areas together provide a practical framework for the responsible deployment of Explainable AI.

A. Problem Statement

Post-hoc numeric explanation methods such as LIME and SHAP produce feature-importance scores that require some degree of statistical literacy to understand. Unfortunately, the vast majority of those most impacted—clinicians, defendants, loan applicants—do not possess this level of statistical literacy. Although LLMs provide an output that is significantly easier to read than post-hoc numeric methods, they also introduce the possibility of confabulation. Each method is imperfect on its own; therefore, getting the combination right requires both good architectural safeguards and proper institutional governance.

B. Contributions

Three contributions are made in this paper: a comparison of LIME, SHAP and LLM on dimensions that are relevant to users who do not have expert knowledge (1); an analysis of the ethics and governance risks that arise due to using language to explain (2); and recommendations for the design and governance of explainable AI systems being used in high-stakes situations. (3).



Fig. 1. XAI layer bridging opaque AI models and end users via natural language explanation.

II. BACKGROUND AND RELATED WORK

A. Public Perception and Governance Motivation

According to survey data, sentiment towards AI has dramatically changed from 2020 to 2023. In 2020, a Pew Research Center survey conducted by themselves with 150K respondents found that AI was viewed positively by 53.75% of all 150K participants. By 2023, two independent studies that were each much larger than 150,000 were done by the Big Data Systems Lab and James with 10,038 respondents, and by Monmouth University with 805 respondents, found that 53% and 41% respectively of the respondents considered AI harmful. This was closely linked to the surge in the media's reporting of algorithmic error and the increasing public awareness of how AI is being used in areas like criminal justice and employment. As a result, regulations have been introduced. The European Union General Data

Protection Regulation [GDPR] provides users with the right to obtain an explanation about automated decisions made about them. The European Union has passed an act that follows up this provision by providing similar means of recourse for users in connection with the use of High Risk A.I. to decision making in healthcare, education, and employment. Further, the United States' National Institute Standards and Technology [NIST] artificial intelligence risk management framework lists explainability as one of the core requirements for establishing trust in artificial intelligence systems. Therefore, the aforementioned regulations establish concrete obligations for technical explainability.

B. Inherently Interpretable Models

Decision trees, logistic regression and linear classifier models—the earliest examples of machine learning—can be interpreted based on their description or appearance. For example, if I see a node on a decision tree, I might think of it as an “if .. then” rule. Likewise, if I see a coefficient representing the effect of one of the variables, I would know that it shows only how much that variable shifts my outcome. However, the reason why these models struggle to perform complex tasks, like natural language understanding and medical imaging, is due to their limitations on capacity and therefore performance. With these challenges in mind, most of the early XAI work attempted to extract interpretable behaviour from more powerful architectures while still maintaining predictions at the same time.

C. Post-Hoc Numeric Explanation: LIME and SHAP

In 2016, Ribeiro, Singh, and Guestrin (2016) presented Local Interpretable Model-Agnostic Explanations (LIME) as a method for explaining the predictions of machine learning models. LIME creates a local explanatory model for each prediction it wants to explain by perturbing an input in a small way, observing the response of the model, and then fitting a simple linear model to the resulting set of responses. The coefficients of the fitted linear model then form the explanation for that prediction. The randomness associated with LIME's random sampling method raises issues with reproducibility. The same input can generate different explanations on different runs—an issue that is particularly difficult to accept in the legal and medical domains (He et al., 2017). Lundberg and Lee (2017) introduced SHAP (SHapley Additive exPlanations) as a method of providing explanations based on the underlying cooperative game theory. SHAP values provide a measure of how much an individual feature contributes, on average, across every possible

combination of the features using the idea of the Shapley value from cooperative game theory to derive SHAP values. SHAP values have three theoretical properties: consistency, local accuracy, and missingness (Lundberg & Lee, 2017). However, calculating exact Shapley values requires an exponential amount of time in relationship to the number of features being assessed, and obtaining an approximate measurement to make computations feasible can require excessively long computation times for particularly large models and data sets (Cohen et al., 2019). According to Markus et al. [8], the main audience of explainable AI (XAI) for clinical applications are clinicians, who often do not have the statistical background required to properly analyse SHAP plots and LIME coefficients. As such, their research on explainability demonstrates the need to balance technical rigor with the practical accessibility of XAI, and transition XAI research into formats of explanation that can be effectively interpreted by non-experts.

D. Natural Language Explanations and LLMs

There has been a significant amount of development in XAI as a result of the emergence of large generative language models (LLMs). Earlier approaches to XAI relied heavily on rules-based templates to translate output from XAI systems into natural language (reference number). Although consistent, they require extensive manual engineering for every new domain and typically feel very mechanical (reference number).Islam et al. [3] and Saeed and Omlin [22] identified language model-based explanations as one of the fastest-growing areas of XAI. Nagahisarchoghaei et al. [23] traced the transition from rule-based NLE systems to learning-based NLE systems using an LLM as a foundation. Zytek et al. [9] proposed the Explingo framework, which uses two LLMs (i.e., the "narrator" and the "grader") to generate and evaluate the quality of narrative explanations produced by SHAP, with strong results in accuracy, completeness, and fluency. Lukassen et al. [24] reported that LLM selection accounts for more variance in NLE quality than either the choice of XAI method or the prompting strategy, and determined that LLM selection is the most important engineering decision for NLE production. Renze [2] provides the best empirical evaluation so far by implementing a self-correcting pipeline of three separate components to evaluate the first 100 criminal risk assessment records produced by the LLMs. The pipeline achieved a factual accuracy of 0.996 for GPT-4 (with self-correction), 0.993 for GPT-4 (without self-correction), and 0.949 for GPT-3.5 (with self-

correction).A hand-coded rule-based baseline hit a ceiling of 1.000, giving an upper bound for comparison.

E. Risks: Confabulation and Fluency Bias

A language model generating an explanation does not have access to a classifier's decision boundary; it constructs a post-hoc narrative from prompt context and training data patterns. When that story is wrong, the polished prose can be more convincing than a dry but accurate numeric output [10], [13]. Fluency bias—the well-documented human habit of treating smooth, confident language as a sign of factual accuracy [25]—is arguably the most serious risk in LLM-based XAI, and it's what drives the quality-assurance architectures we review in Section V.

TABLE I

SUMMARY OF KEY XAI STUDIES

Author(s)	Year	Contribution	Key Result	Limitation
Ribeiro et al.	2016	LIME: local surrogate	Model-agnostic XAI	Unstable across runs
Lundberg & Lee	2017	SHAP: Shapley values	Axiomatic guarantees	High compute cost
Markus et al.	2021	Healthcare XAI review	Clinicians need NL output	Numeric methods fail
Islam et al.	2021	XAI survey	Catalogued 50+ methods	Limited LLM focus
Saeed & Omlin	2023	Meta-survey of XAI gaps	Usability gaps flagged	Needs richer formats
Zytek et al.	2024	Explingo NLE system	High accuracy & fluency	Limited domains tested
Renze	2023	GPT self-correcting NLE	GPT-4 accuracy = 0.996	Single dataset only



Fig. 2. Comparison of LIME, SHAP, and LLM-based approaches on key explainability dimensions.

III. GOVERNANCE STRATEGIES FOR SAFE AI

James [1] argues that technical explainability is a prerequisite for responsible AI, but it is not even close to being sufficient. The questions of who builds these systems, what they may learn, and who ultimately authorizes their deployment require institutional answers that no explanation architecture can provide. This section walks through the governance strategies to be laid out in [1].

A. Limiting AI Learning Scope

AI training and development using unfiltered online data has resulted in large amounts of untrue, inaccurate, and harmful information being fed into models. According to James [1], there is an analysis of online data that indicates approximately 62% cannot be trusted as accurate. In addition, social media content-based algorithms exacerbate this issue by steering users towards content that matches their pre-existing beliefs or biases. By limiting training to sources that have been reviewed/certified, we can reduce the amount of false information getting into AI generated outcomes [1].

B. Avoiding Centralised AI Architectures

An AI system dominating all other AI systems and using all their outputs would create a concentration of capability that would be difficult to audit or control. James advocates limiting AI systems to narrowly defined tasks to facilitate auditing and reduce the chance of unexpected behaviour across different areas. This is consistent with the modular structure of the Explainer-Verifier-Corrector pipeline assessed in [2].

C. Central Committee for Model Approval

An independent review organization could be used to examine (as a model) before it is released to the public, similarly to institutional review boards that exist for the purpose of overseeing research in medicine [1]. This organization would evaluate both the anticipated use of

the classification system and its potential to be misused, and as a result, this organization would have actual authority to block the release of the model. This proposal would be consistent with the requirement for high-risk AI systems to complete the evaluation of conformity with the EU AI Act, outlined in the High-Risk AI System – Conformity Assessment Document [16].

D. Ethical Implications of AI Weapons

According to James [1], an area in which we might lose control over the use of autonomy is in the creation of autonomous weapons systems which produce effects that cannot be undone. The deployment of lethal autonomous systems without keeping humans involved in any part of the decision-making process would be a violation of the principle of meaningful human oversight that is considered fundamental to all major governance frameworks [17].

IV. METHODOLOGY

A. Research Design

This is a structured synthesis of literature using empirical studies in the form of primary sources [1],[2] and secondary study of XAI studies in literature [3] – [25]. The difference between a narrative review & structured literature synthesis is that the structured synthesis defines in advance what the inclusion/exclusion criteria will be for study materials, how the literature search was performed, and how each empirical study was analyzed with the same lens [22].

B. Primary Empirical Sources

There are two main sources used in this research paper. James [1] provided data about people's perceptions through three separate studies involving over 160,000 participants, as well as a set of proposed governance structures. Renze [2] provided quantitative evaluation results for four different GPT models based on 100 COMPAS Synthetic case records (including accuracy, runtime and cost metrics). Both datasets have been made accessible to the public.

C. Literature Search

Through an extensive literature search using the following search terms: "explainable artificial intelligence," "XAI," "natural language explanations," "large language models," "LIME," "SHAP," "governance of AI," "human oversight," "confabulation," and "trustworthy AI" in the databases IEEE Xplore, ACM Digital Library, arXiv, and Google Scholar, secondary sources were identified. All publications were limited to

peer-reviewed articles published from 2016 to the present. A total of 25 sources were included after screening for relevance and quality.

D. Inclusion and Exclusion Criteria

Peer-reviewed articles focusing on AI governance or Explainable Artificial Intelligence (XAI) methods constitute eligible papers for inclusion into this research. Papers must be available in English with adequate detail regarding methodology; otherwise, they are excluded if: only touching on interpretability, dealing only with computer vision while omitting generalisation issues; non-empirical position pieces without analytic support.



Fig. 3. Research methodology: from question formulation through structured synthesis.

V. LLM-BASED EXPLANATION ARCHITECTURE

A. System Components

The Renze [2] evaluation used three components of a GPT-type system in series: An Explainer that received the formatted case record data (key value pairs of feature scores with their respective SHAP importance scores) together with a system prompt and some example records to generate a letter of natural language explanation; a Verifier that checked the generated letter for factual accuracy against the case record; and a Corrector that repaired any errors identified from the output. Prior to sending features into the LLM, the features were transformed and scaled to be in-range of 0–1 and categorically labelled low, medium or high. This labelling step proved to be quite significant, as earlier experimentation providing the raw SHAP values were appreciably less successful in performance than those providing categorical representations. Consequently, the results imply that categorical representations are likely to be better suited for LLM processing than providing raw numerical inputs. [2].

B. Self-Correction Mechanism

GPT-4 has demonstrated an accuracy of 0.993 when it assesses facts with no self-correction. There are errors in two places: incorrect descriptive level (low, medium, or high) for a given feature score, and incorrect placement of features in the paragraphs based on their significance. The Verifier identifies these errors by using a ground-truth scoring rubric that compares the claims made in explanations (in all forms) against those claims' respective ground truths. Once the Verifier identifies an error as described above, the Corrector will create an entirely new version of the passage to increase accuracy from 0.993 to 0.996 [2].



Fig. 4. Explainer-Verifier-Corrector pipeline with feedback loop for iterative refinement [2].

C. Comparison with Rule-Based Systems

The design of a rule-based generator (hand-coded) reached an accuracy threshold of 1.000 after developing human-written templates and specifying explicit rules for each grammar exception (e.g., pluralisation, conjunctions, blank paragraphs). The GPT method achieved comparable quality without having to develop domain-specific programs and brought with it features that no rule-based generator could match: an ability to adapt reading levels to suit an audience; support for different languages; and an ability to answer follow-up queries.

D. Fluency Bias and Mitigation

The fluency of LLM (late language model) explanations is their most significant strength and their most serious problem. Studies have shown that individuals view fluent writing as being more credible, regardless of its accuracy compared to a bare list of numerical statistics [10], [13], [25]. To help mitigate this risk, it is recommended that explicit statements expressing uncertainty accompany the generated explanations; having a defined minimum threshold for quality so only outputs above this minimum groundedness score would be accepted; conducting A/B tests with non-expert users to determine if there was an

actual improvement in decision making quality versus only measuring the accuracy of the explanation [9], [24].

VI. RESULTS AND COMPARATIVE ANALYSIS

A. Quantitative Performance of GPT Models

According to Renze [2], Table II compares four different GPT configurations against the ceiling set by rule-based systems. The configuration of GPT with a self-correction capability (GPT-4) produced an output with a factual accuracy rating of 0.996, and its explanatory corrections cost \$0.039 each (which is less than the cost of \$0.054 for the explanation). The configuration of GPT-3 with self-correction produced an output with a factual accuracy rating of 0.949 but was approximately 16 times cheaper and faster than the self-correction capabilities of GPT-3.

TABLE II

GPT MODEL PERFORMANCE SUMMARY (RENZE, 2023 [2])

Model	Accuracy	Runtime/expl.	Cost/expl.	Suitable for?
Rules baseline	1.000	< 1 s	~\$0.000	Any domain (rigid)
GPT-4 corrected	0.996	37.8 s	\$0.039	High-stakes (legal, medical)
GPT-4 baseline	0.993	35.8 s	\$0.097	Moderate-stakes
GPT-3.5 corrected	0.949	21.0 s	\$0.007	Low-stakes / high volume
GPT-3.5 baseline	0.938	15.9 s	\$0.006	Exploratory only

B. Survey Data: Shifting Public Attitudes

The data from James [1] is included in Table III. The trend across independent surveys with diverse methods shows a decrease in positive views on AI from the majority to the minority, while negative views have increased significantly. These findings support the empirical basis for the governance recommendations presented in Section III.

TABLE III

PUBLIC PERCEPTION OF AI: SURVEY EVIDENCE (JAMES, 2023 [1])

Source	Year	n	Positive	Negative / Concerned
Pew Research Center	2020	150,000	53.75%	32.45%
Big Data System Lab & James	2023	10,038	35%	53%
Monmouth University	2023	805	9%	41%

C. Comparison of Explanation Approaches

From analyzing both the data presented in Tables I and II, in addition to reviewing the broader literature in general, three key conclusions can be drawn. First, LLM-generated explanations exhibit significantly greater flexibility and language proficiency than rule-based systems; however, they are slightly less factually accurate overall. Second, there is sufficient disparity in accuracy between GPT-3.5 (0.938) and GPT-4 (0.993) that it could impact decision making in higher-risk circumstances while still being reasonable enough to allow continued use in lower-risk situations. Thirdly, implementing self-correcting features provides measurable gains in accuracy (i.e., 0.003 additional accuracy with GPT-4; 0.011 additional accuracy with GPT-3.5) when compared to using them without self-correction; therefore, this represents an acceptable and valuable investment for any production systems [2],[9].

VII. DISCUSSION

A. Complementarity of Technical and Governance Approaches

The two articles, James [1] regarding governance strategy and Renze [2] regarding technical architecture, address the same core problem through different layers of that problem. Governance addresses systemic risk—who creates AI, how does AI learn, and under whom does AI operate. Technical explainability addresses the individual decision-making process—that is, how a specific output will be communicated to its affected party. Both of these elements are absolutely critical to successful governance

of AI. A fully transparent environment without governance oversight creates risk; an environment that has governance oversight, but cannot demonstrate its technical decision-making capability, removes the opportunity for real human control.

B. The Non-Expert Gap

James's [1] research on survey responses from different countries provides clear evidence that individuals are not concerned about AI primarily as a statistical concept, but rather due to high-profile cases where AI has failed users, and because of their own experiences using an algorithmic decision process where they were unable to comprehend or challenge it. In their analysis of how to provide non-experts with valid information about AI decision making, Renze [2], and Zytek et al. [9], determined that natural language explanation is currently the best means available to provide ways for non-experts to gain insight into how AI makes its decisions. These may contain errors, which is the purpose of the Self-Correcting Architecture—an application designed to prevent this from happening [2].

C. Regulatory Alignment

Meaningful explanations must accompany consequential automated decisions as required by the GDPR [15], the EU AI Act [16], and the NIST AI RMF [17]. The combination surveyed here (governance committees, scoped training, and self-correcting NLE pipelines) provides a clear path toward compliance with regulations based on both technological and institutional principles. However, compliance should be considered the minimum requirement for governing AI; compliance is not enough—AI should, first and foremost, serve human needs, rather than simply meeting regulatory requirements.

VIII. FUTURE RESEARCH DIRECTIONS

1) To determine the extent to which the automated correction pipeline is generalizable across domains, replicate Renze's [2] experiment using standard XAI benchmark datasets. The standard datasets used for Renze's research include the COMPAS recidivism prediction dataset, the 1994 Adult Income census survey, and the German Credit data set.

2) Conduct empirical user studies that measure the extent to which LLM-generated explanations influence the performance of non-expert decision-makers compared to simply assessing users' perceived quality of explanations [9, [24]].

3) Create specific, domain-based rubrics that define what constitutes an adequate explanation for medical clinical decisions, financial decisions, and legal decisions, based on existing clinical AI guidelines and the work of Markus et al. [8].

4) Conduct empirical evaluations of different AI governance intervention options, e.g. to determine whether AI misuse has decreased through the establishment of model-approval committees and curation of training data.

5) Create new methods / combine methods by chaining LIME/SHAP to create an attribution value for a prediction, and narrate that attribution using LLM, allowing the Verifier to check that the explanation provided is in agreement with the numerical ground truth provided in the (Attribution) dataset.

6) Actively measure the fluency bias (i.e. the perceived quality of a written communication as a function of the writer's linguistic style), and to isolate any changes in users' level of acceptance of LLM-generated explanations, regardless of their accuracy, from the findings in [10, [13], [25]] on the relationship between the quality of an explanation and user satisfaction.

IX. CONCLUSION

The previous article discussed the concept of explainability through AI models from two angles - the technical aspect of implementing models that can be easily read by non-experts using natural language generation, and the governance aspect of controlling AI so that it remains under the mandate of genuine human control, both of which are important, but together do not completely solve the problem. Technically speaking, Renze (2018) achieved 0.996 as the factual accuracy of the self-correcting GPT pipeline, which nearly meets the accuracy of a rule-based hand-coded system, while still providing a more flexible and natural experience. Governance-wise, James (2020) conducts a survey which provides evidence of a shift in public sentiment, and as such, James identified four specific governance interventions; scoped training, decentralized architecture, model-approval committees, and organizations providing human-in-the-loop requirements for high stakes uses. The literature in the field of explainable AI evaluated in this article indicates that the two threads discussed are indeed converging, as there seems to be a consensus in the literature that the provision of explanation through LLM-based models will be the most promising means of reaching explainable AI, and the noted governance frameworks will become binding law. The combination

of researchers/practitioners who are able to bridge both threads and build systems that cover both aspects has the best opportunity to successfully implement AI that is truly explainable, truly ethical, and truly trustworthy.

REFERENCES

- A. J. James, "Human oversight and control in AI development and deployment," *Technology in Society*, Preprint submitted May 31, 2023.
- M. Renze, "Natural language explanations for explainable AI systems using large language models with self-correcting architectures," *arXiv preprint*, Johns Hopkins University, 2023.
- S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," *arXiv:2101.09429*, 2021.
- M. Nagahisarchoghaei et al., "An empirical survey on explainable AI technologies: Recent trends, use-cases, and categories from technical and application perspectives," *Electronics*, vol. 12, no. 5, p. 1092, 2023.
- J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, 2016.
- S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Computers in Biology and Medicine*, vol. 156, 2023.
- A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care," *Journal of Biomedical Informatics*, vol. 113, 2021.
- A. Zytek, S. Pidò, S. Alnegheimish, L. Berti-Équille, and K. Veeramachaneni, "Explingo: Explaining AI predictions using large language models," *arXiv:2412.05145*, 2024.
- P. Biecek et al., "Explainable AI (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions," *Information Fusion*, vol. 106, 2024.
- S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
- M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 1135–1144, 2016.
- Creative Solutions Coaching, "When AI sounds right: How fluency creates false confidence in decision-making," *creativesolutionscoaching.com*, Jan. 2026.
- R. Tomsett et al., "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems," in *Proc. ICML Workshop on Human Interpretability in Machine Learning*, 2018.
- European Parliament, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Official Journal of the European Union, 2016.
- European Parliament, "Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act)," Official Journal of the European Union, 2024.
- National Institute of Standards and Technology, "AI Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023.
- D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in *Proc. ICML Workshop on Human Interpretability in Machine Learning*, 2018.
- V. V. Strumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, 2014.
- I. Donadello and M. Dragoni, "Bridging signals to natural language explanations with explanation graphs," in *Proc. 2nd Italian Workshop on Explainable AI*, 2021, pp. 1–13.
- IBM, "AI Explainability 360," IBM Research Trusted AI, *aix360.mybluemix.net*, 2023.
- W. Saeed and C. W. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, 2023.
- M. Nagahisarchoghaei et al., "Explainable AI in practice: Taxonomies, challenges, and recent advances," *Electronics*, 2023.
- F. Lukassen et al., "From XAI to stories: A factorial study of LLM-generated explanation quality," *arXiv:2601.02224*, 2026.
- R. Hoffman et al., "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance," *Frontiers in Computer Science*, vol. 5, 2023.