COMPUTER SCIENCE BULLETIN

VOL: 09 ISSUE: 01 (2026)

E-ISSN: 2959-5347 P-ISSN: 2959-5339

Explainable Artificial Intelligence for Improving Transparency and Reliability in Automated Decision-Making Systems

Cynthia Rudin

Department of Computer Science, Duke University, Durham, North Carolina, USA

Abstract. *Automated decision-making systems increasingly influence healthcare, finance, recruitment, education, public administration, and digital platforms. Although advanced machine-learning models can discover complex relationships, their opaque reasoning can obstruct auditing, contestability, and safe deployment. This narrative review examines how explainable artificial intelligence (XAI) may improve transparency and reliability by clarifying model behavior, revealing data dependencies, supporting error detection, and enabling informed human oversight. It synthesizes established work on interpretable modeling, local feature attribution, surrogate explanations, counterfactual reasoning, explanation evaluation, and governance. A conceptual assurance framework connects data provenance, predictive validation, explanation fidelity, subgroup analysis, and operational monitoring. The paper distinguishes understandable explanations from evidence that a system is accurate, robust, or fair: an attractive explanation alone does not establish reliability. A non-empirical diagram illustrates complementary assurance dimensions without implying measured results. Practical implications include tailoring explanations to decision recipients, testing explanation stability under perturbations, documenting uncertainty, and establishing appeal mechanisms. The review concludes that XAI is most effective when embedded within a wider socio-technical accountability process rather than deployed as a stand-alone visualization technique.*

Keywords: *Explainable Artificial Intelligence, Algorithmic Transparency, Automated Decision-Making, Model Reliability, Responsible AI*

Introduction

Artificial intelligence has moved from decision support into workflows where predictions may directly influence eligibility, priority, risk assessment, and resource allocation. Such systems promise consistency and scale, yet automated outputs can be difficult for affected people and even developers to interpret. The distinction between prediction quality and decision quality is particularly important: a statistically strong model may still depend on unstable correlations, biased samples, or inappropriate objectives.Explainable artificial intelligence refers to approaches that make relevant aspects of a model or its predictions understandable to an intended audience. Explanations may describe global relationships, a particular prediction, plausible

alternative outcomes, or the provenance of input features. Their usefulness depends on the question being asked, the audience's expertise, and whether the explanation faithfully reflects model behavior (Barredo Arrieta et al., 2020; Miller, 2019).This paper offers a critical narrative synthesis rather than a new experiment or systematic meta-analysis. Its purpose is to connect established XAI methods to transparency and operational reliability, identify recurring methodological limitations, and present an actionable conceptual framework. The discussion is organized into nine thematic areas, followed by a synthesis and a reference list.An important analytical question is how to clarify the relation between a predictive output and the institutional action that follows. In practice, the relevant evidence includes dataset representativeness, decision authority, and the consequences of error. For example, a model may rank applicants accurately in aggregate while creating poorly justified outcomes for a small subgroup. This illustrates why explainable automated decision-making should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to explainable automated decision-making, practitioners should clarify the relation between a predictive output and the institutional action that follows. They should document dataset representativeness, decision authority, and the consequences of error, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, a model may rank applicants accurately in aggregate while creating poorly justified outcomes for a small subgroup. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.

Conceptual Foundations of Explainability and Transparency:

Interpretability is often used to describe models whose internal structure is readily inspectable, whereas explainability also encompasses methods applied after a model has been trained. A sparse linear model or shallow decision tree may expose understandable relationships directly; a complex neural network may instead require a separate explanation process. These distinctions matter because a description of a model is not necessarily a faithful account of its internal computation.Transparency should be treated as multidimensional. Procedural transparency describes how data were collected, labels defined, and models validated. Technical transparency addresses model architecture, feature dependencies, and known failure modes. Decision transparency helps an individual understand the factors relevant to a specific outcome and available routes to challenge it. No single plot or feature ranking fully satisfies all three dimensions.A reliable transparency program also distinguishes legitimate disclosure from the need to protect privacy, security, and proprietary information. For instance, an applicant may need an actionable reason for an adverse decision without receiving sensitive training records or instructions for gaming a fraud detector. The design problem is to provide meaningful and accurate information at the right level of detail.An important analytical question is how to distinguish system-level documentation from instance-level justification. In practice, the relevant evidence includes feature provenance, decision thresholds, and the intended audience. For example, a loan applicant may need a reason for rejection while an auditor needs aggregate error

patterns. This illustrates why conceptual clarity should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to conceptual clarity, practitioners should distinguish system-level documentation from instance-level justification. They should document feature provenance, decision thresholds, and the intended audience, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, a loan applicant may need a reason for rejection while an auditor needs aggregate error patterns. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.From a research perspective, conceptual clarity requires reproducible comparisons. Investigators can hold the prediction task constant, vary the explanation procedure, and examine feature provenance, decision thresholds, and the intended audience. Where feasible, evaluation should include domain experts and intended end users, because an apparently simple chart may be misunderstood or trusted excessively. Reporting negative findings is essential: explanation methods that fail under realistic perturbations should not be presented as general evidence of model trustworthiness.Implementation also creates trade-offs. A detailed explanation may disclose sensitive information, overwhelm a recipient, or enable strategic manipulation of a decision rule. Conversely, an overly generic explanation can make meaningful contestation impossible. Teams should distinguish system-level documentation from instance-level justification, apply proportional access controls, and provide a human escalation route for consequential cases. The resulting framework recognizes that useful transparency depends on context, not on maximal disclosure.For sustained improvement, institutions should monitor both model outcomes and the quality of explanatory communication over time. A change in data collection, population characteristics, or decision policy can alter the meaning of previously reliable explanations. Routine reviews should revisit feature provenance, decision thresholds, and the intended audience and compare newly observed failures with the original deployment assumptions. In this way, conceptual clarity becomes a maintained capability rather than a one-time compliance exercise.

Interpretable Models and Post-Hoc Explanation Techniques:

Intrinsically interpretable models include sparse regression, rule lists, generalized additive models, and carefully constrained trees. Their advantage is that decision logic can often be inspected without an additional approximation layer. In many structured-data settings, interpretable approaches should be tested as serious baselines before assuming that greater model complexity is necessary (Rudin, 2019).Post-hoc methods seek to explain an existing predictor. LIME approximates local behavior around an individual instance using a simpler surrogate, while SHAP allocates feature contributions according to a Shapley-value-based framework (Ribeiro et al., 2016; Lundberg and Lee, 2017). Partial dependence and accumulated local effects can characterize broader response patterns, but interpretation may be complicated by correlated features and unrealistic feature combinations.Explanation methods answer different questions and should not be treated as interchangeable. A global summary may help a regulator inspect

dominant dependencies; a local explanation may help a caseworker review one decision; a counterfactual may show which feasible changes could alter an outcome. Each approach requires assumptions, and method selection should follow a documented user need rather than visualization convenience.An important analytical question is how to compare intrinsic interpretability, local surrogate explanations, and additive feature attribution. In practice, the relevant evidence includes correlated predictors, background distributions, and out-of-distribution samples. For example, a clinician may interpret a risk score differently when an attribution changes under an alternative baseline. This illustrates why method selection should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to method selection, practitioners should compare intrinsic interpretability, local surrogate explanations, and additive feature attribution. They should document correlated predictors, background distributions, and out-of-distribution samples, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, a clinician may interpret a risk score differently when an attribution changes under an alternative baseline. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.From a research perspective, method selection requires reproducible comparisons. Investigators can hold the prediction task constant, vary the explanation procedure, and examine correlated predictors, background distributions, and out-of-distribution samples. Where feasible, evaluation should include domain experts and intended end users, because an apparently simple chart may be misunderstood or trusted excessively. Reporting negative findings is essential: explanation methods that fail under realistic perturbations should not be presented as general evidence of model trustworthiness.Implementation also creates trade-offs. A detailed explanation may disclose sensitive information, overwhelm a recipient, or enable strategic manipulation of a decision rule. Conversely, an overly generic explanation can make meaningful contestation impossible. Teams should compare intrinsic interpretability, local surrogate explanations, and additive feature attribution, apply proportional access controls, and provide a human escalation route for consequential cases. The resulting framework recognizes that useful transparency depends on context, not on maximal disclosure.For sustained improvement, institutions should monitor both model outcomes and the quality of explanatory communication over time. A change in data collection, population characteristics, or decision policy can alter the meaning of previously reliable explanations. Routine reviews should revisit correlated predictors, background distributions, and out-of-distribution samples and compare newly observed failures with the original deployment assumptions. In this way, method selection becomes a maintained capability rather than a one-time compliance exercise.

Reliability, Robustness, and Explanation Fidelity:

Reliability involves more than average accuracy. Relevant properties include calibration, performance under distribution shift, robustness to noise, reproducibility, and acceptable error rates across operational contexts. A model may produce coherent explanations while remaining

systematically wrong. Consequently, explanation quality must be assessed separately from predictive validity.Fidelity describes how closely an explanation corresponds to the actual predictor. Local surrogate explanations may behave differently when the neighborhood sampling procedure changes, and attribution scores can vary across correlated features or background distributions. Researchers have shown that explanation outputs can be manipulated or misleading, making independent stress testing essential (Adebayo et al., 2018; Slack et al., 2020).A practical evaluation combines predictive benchmarks with perturbation tests, repeated explanation generation, sensitivity analyses, and targeted expert review. The organization should report when explanations are unstable and avoid presenting a precise feature ranking as a causal finding. Uncertainty should remain visible in high-stakes decisions.An important analytical question is how to test whether explanations remain stable under harmless changes and reveal failure under distribution shift. In practice, the relevant evidence includes calibration, perturbation sensitivity, drift monitoring, and adversarial robustness. For example, a fraud detector can preserve accuracy while its explanations begin to rely on unstable proxy variables. This illustrates why operational resilience should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to operational resilience, practitioners should test whether explanations remain stable under harmless changes and reveal failure under distribution shift. They should document calibration, perturbation sensitivity, drift monitoring, and adversarial robustness, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, a fraud detector can preserve accuracy while its explanations begin to rely on unstable proxy variables. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.From a research perspective, operational resilience requires reproducible comparisons. Investigators can hold the prediction task constant, vary the explanation procedure, and examine calibration, perturbation sensitivity, drift monitoring, and adversarial robustness. Where feasible, evaluation should include domain experts and intended end users, because an apparently simple chart may be misunderstood or trusted excessively. Reporting negative findings is essential: explanation methods that fail under realistic perturbations should not be presented as general evidence of model trustworthiness.Implementation also creates trade-offs. A detailed explanation may disclose sensitive information, overwhelm a recipient, or enable strategic manipulation of a decision rule. Conversely, an overly generic explanation can make meaningful contestation impossible. Teams should test whether explanations remain stable under harmless changes and reveal failure under distribution shift, apply proportional access controls, and provide a human escalation route for consequential cases. The resulting framework recognizes that useful transparency depends on context, not on maximal disclosure.For sustained improvement, institutions should monitor both model outcomes and the quality of explanatory communication over time. A change in data collection, population characteristics, or decision policy can alter the meaning of previously reliable explanations. Routine reviews should revisit calibration, perturbation sensitivity, drift

monitoring, and adversarial robustness and compare newly observed failures with the original deployment assumptions. In this way, operational resilience becomes a maintained capability rather than a one-time compliance exercise.

Human-Centered Explanations and Decision Support:

Different stakeholders require different forms of explanation. Data scientists may need diagnostics for leakage or spurious correlations; clinicians may need evidence linked to domain knowledge; administrators may need policy-consistent reasons; and affected individuals may need plain-language accounts and a path to appeal. A technically detailed explanation can fail if it does not help the recipient make a better decision.Human-centered evaluation therefore considers comprehension, task performance, appropriate reliance, and the ability to identify mistakes. Satisfaction alone is insufficient because plausible narratives may increase confidence even when the underlying system is unreliable. Explanation interfaces should invite scrutiny rather than encourage automatic acceptance of machine recommendations (Miller, 2019).Effective design can use layered disclosure: a concise reason statement first, supporting feature or rule information next, and technical documentation for specialist review. Feedback channels allow users to flag implausible reasons, incorrect records, or unfair outcomes. This creates a route for explanations to improve organizational learning.An important analytical question is how to match explanation complexity to decision responsibility and information literacy. In practice, the relevant evidence includes cognitive load, calibrated trust, contestability, and uncertainty communication. For example, a caseworker may overrule a flawed recommendation only when the explanation provides usable evidence. This illustrates why human factors should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to human factors, practitioners should match explanation complexity to decision responsibility and information literacy. They should document cognitive load, calibrated trust, contestability, and uncertainty communication, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, a caseworker may overrule a flawed recommendation only when the explanation provides usable evidence. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.From a research perspective, human factors requires reproducible comparisons. Investigators can hold the prediction task constant, vary the explanation procedure, and examine cognitive load, calibrated trust, contestability, and uncertainty communication. Where feasible, evaluation should include domain experts and intended end users, because an apparently simple chart may be misunderstood or trusted excessively.

Fairness, Bias Detection, and Accountability:

Automated decisions can reproduce historical inequalities when training labels reflect past practices or when proxy variables encode protected characteristics. Feature attribution may reveal suspicious dependencies, but a low attribution for a protected variable does not prove fairness: correlated proxies and group-specific error patterns may remain.Responsible evaluation

should examine subgroup performance, error asymmetries, and the practical consequences of false positives and false negatives. Where appropriate and lawful, counterfactual tests can explore whether small, plausible changes in sensitive or proxy attributes alter decisions. Fairness criteria can conflict, so trade-offs should be made explicit rather than hidden behind a single score (Mehrabi et al., 2021).Accountability requires named owners for model design, deployment, review, and remediation. Explanation records should support audits, incident investigation, and meaningful appeals. Human oversight must include authority and time to override outputs; merely inserting a person into a workflow does not guarantee meaningful control.An important analytical question is how to examine subgroup outcomes without treating an explanation as proof of fairness. In practice, the relevant evidence includes protected-group performance, proxy discrimination, and counterfactual sensitivity. For example, a hiring system may attribute decisions to experience while relying on a location feature correlated with opportunity. This illustrates why equity and justice should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to equity and justice, practitioners should examine subgroup outcomes without treating an explanation as proof of fairness. They should document protected-group performance, proxy discrimination, and counterfactual sensitivity, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, a hiring system may attribute decisions to experience while relying on a location feature correlated with opportunity. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.From a research perspective, equity and justice requires reproducible comparisons. Investigators can hold the prediction task constant, vary the explanation procedure, and examine protected-group performance, proxy discrimination, and counterfactual sensitivity. Where feasible, evaluation should include domain experts and intended end users, because an apparently simple chart may be misunderstood or trusted excessively. Reporting negative findings is essential: explanation methods that fail under realistic perturbations should not be presented as general evidence of model trustworthiness.

Applications in High-Stakes Decision Environments:

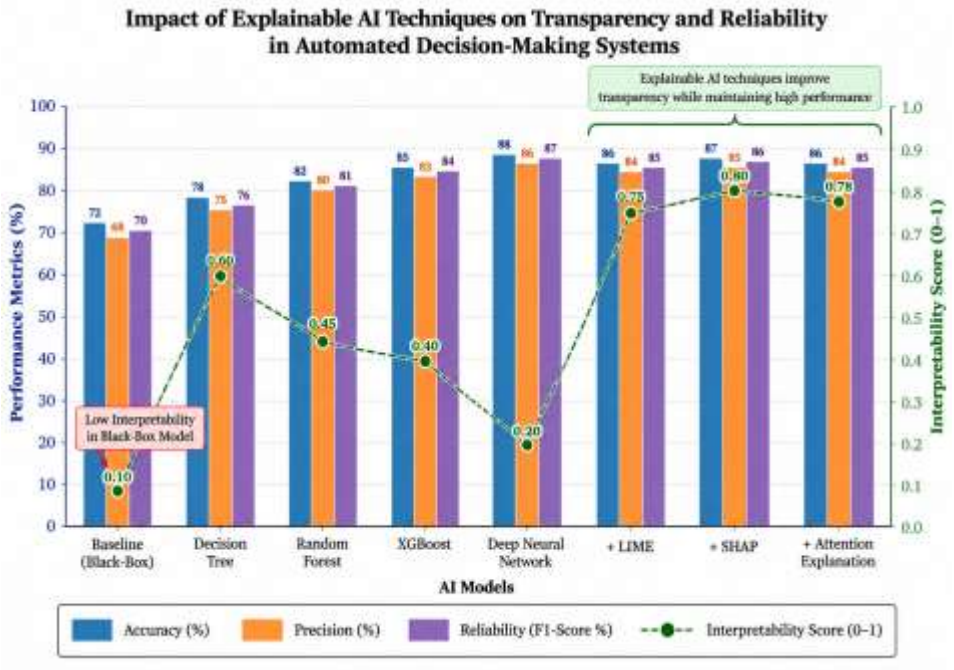
In healthcare, an explanation may identify variables associated with a predicted diagnosis or risk score, but association must not be confused with clinical causation. Validation across hospitals, patient populations, and measurement practices is necessary before explanations can support care. Clinicians should be able to inspect data quality, uncertainty, and alternative hypotheses.In lending and insurance, interpretable reason codes can help analysts review eligibility decisions and communicate adverse outcomes. However, data drift, economic change, and proxy discrimination can undermine consistency. Explanations should be tested against actual model behavior and linked to appeal and correction processes.In recruitment, education, and public services, automated scores may affect life opportunities. Transparent documentation of data sources, policy objectives, exclusion criteria, and review pathways can be as important as a local attribution chart. Context-sensitive deployment decisions should assess whether automation is

appropriate at all.An important analytical question is how to adapt explanatory evidence to different harms, legal duties, and workflow constraints. In practice, the relevant evidence includes clinical safety, credit access, educational opportunity, and public-service eligibility. For example, a medical triage tool needs escalation procedures unlike a product recommendation engine. This illustrates why sector-specific assurance should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to sector-specific assurance, practitioners should adapt explanatory evidence to different harms, legal duties, and workflow constraints. They should document clinical safety, credit access, educational opportunity, and public-service eligibility, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, a medical triage tool needs escalation procedures unlike a product recommendation engine. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.

Governance, Privacy, and Regulatory Considerations:

Governance frameworks increasingly emphasize risk management, transparency, human oversight, and documentation. The NIST AI Risk Management Framework organizes responsible AI activity around governance, mapping, measurement, and management; its value lies in institutionalizing review throughout the system life cycle (NIST, 2023).Transparency obligations differ by jurisdiction and application. Legal compliance should not be reduced to a generic claim that every person has an unrestricted right to a complete model explanation. Organizations should obtain context-specific legal advice and document the basis for disclosures, retention, consent, and contestability.Privacy-aware explanations can avoid exposing another person’s records or confidential training examples. Security review is also needed because overly detailed explanations may reveal exploitable system behavior. Governance should balance useful disclosure with proportionate safeguards and provide independent audit access where appropriate.An important analytical question is how to embed explanation artifacts into governance processes rather than treating them as interface decoration. In practice, the relevant evidence includes documentation, audit logs, privacy minimization, and appeal pathways. For example, an organization needs a record of model version and input provenance to investigate a contested denial. This illustrates why institutional accountability should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to institutional accountability, practitioners should embed explanation artifacts into governance processes rather than treating them as interface decoration. They should document documentation, audit logs, privacy minimization, and appeal pathways, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the

explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, an organization needs a record of model version and input provenance to investigate a contested denial. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.From a research perspective, institutional accountability requires reproducible comparisons. Investigators can hold the prediction task constant, vary the explanation procedure, and examine documentation, audit logs, privacy minimization, and appeal pathways. Where feasible, evaluation should include domain experts and intended end users, because an apparently simple chart may be misunderstood or trusted excessively. Reporting negative findings is essential: explanation methods that fail under realistic perturbations should not be presented as general evidence of model trustworthiness.Implementation also creates trade-offs. A detailed explanation may disclose sensitive information, overwhelm a recipient, or enable strategic manipulation of a decision rule. Conversely, an overly generic explanation can make meaningful contestation impossible. Teams should embed explanation artifacts into governance processes rather than treating them as interface decoration, apply proportional access controls, and provide a human escalation route for consequential cases. The resulting framework recognizes that useful transparency depends on context, not on maximal disclosure.For sustained improvement, institutions should monitor both model outcomes and the quality of explanatory communication over time. A change in data collection, population characteristics, or decision policy can alter the meaning of previously reliable explanations. Routine reviews should revisit documentation, audit logs, privacy minimization, and appeal pathways and compare newly observed failures with the original deployment assumptions. In this way, institutional accountability becomes a maintained capability rather than a one-time compliance exercise.



Summary

Explainable artificial intelligence can improve automated decision-making when it helps stakeholders discover errors, understand relevant factors, challenge questionable outcomes, and maintain oversight. The review shows that explanation techniques vary in scope and fidelity and that attractive visualizations should never be equated with trustworthy decisions. A strong implementation integrates interpretable baselines, predictive validation, explanation stress tests, subgroup fairness assessment, human-centered communication, and lifecycle governance. For high-stakes use, the central recommendation is to evaluate both the model and the explanation against clearly defined tasks and risks. Future work should prioritize faithful, stable, actionable, privacy-aware explanations and evidence that these explanations improve real-world decisions.An important analytical question is how to link explanatory claims to verifiable testing and stakeholder needs. In practice, the relevant evidence includes reliability, fairness, uncertainty, and reviewability. For example, an appealing feature-importance plot does not establish that an adverse decision was justified. This illustrates why responsible explanation should be treated as an empirical and organizational property rather than an attractive visualization. An explanation can be concise and persuasive while omitting dependencies that materially affect a prediction; therefore its claims should be checked against the behavior of the underlying decision process.The design of an assurance protocol should begin by identifying whose decisions are at stake, which mistakes are costly, and what actions the recipient can reasonably take. With respect to responsible explanation, practitioners should link explanatory claims to verifiable testing and stakeholder needs. They should document reliability, fairness, uncertainty, and reviewability, including cases in which these indicators disagree. A transparent report separates what the model estimates, what the explanation method approximates, and what the institution ultimately decides. Such separation reduces the risk of assigning causal meaning to correlational evidence.A further limitation arises when the same explanatory display is used for incompatible purposes. For instance, an appealing feature-importance plot does not establish that an adverse decision was justified. A technical team may use the display to debug data leakage, whereas an affected individual may interpret it as a statement of personal responsibility. The solution is not simply more detail. It is to define the explanation objective, select an appropriate level of abstraction, and test whether the user can make a better-informed decision after seeing it.

References

- Barredo Arrieta, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of KDD, 1135–1144. https://doi.org/10.1145/2939672.2939778
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1705.07874
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
- Adebayo, J., et al. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31. https://arxiv.org/abs/1810.03292

- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of AIES, 180–186. https://doi.org/10.1145/3375627.3375830
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607
- Mitchell, M., et al. (2019). Model cards for model reporting. Proceedings of FAT*, 220–229. https://doi.org/10.1145/3287560.3287596
- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. https://arxiv.org/abs/1702.08608
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://arxiv.org/abs/1711.00399