Bridging the Gap in XAI-Why Reliable Metrics Matter for Explainability and Compliance

Pratinav Seth
Lexsi Labs
Mumbai, India
pratinav.seth@lexsi.ai

Vinay K. Sankarapu
Lexsi Labs
London, United Kingdom
v.k@lexsi.ai

Abstract

Reliable explainability is not only a technical goal but also a cornerstone of private AI governance. As AI models enter high-stakes sectors, private actors such as auditors, insurers, certification bodies, and procurement agencies require standardized evaluation metrics to assess trustworthiness. However, current XAI evaluation metrics remain fragmented and prone to manipulation, which undermines accountability and compliance. We argue that standardized metrics can function as *governance primitives*, embedding auditability and accountability within AI systems for effective private oversight. Building upon prior work in XAI benchmarking, we identify key limitations in ensuring faithfulness, tamper resistance, and regulatory alignment. Furthermore, interpretability can directly support model alignment by providing a verifiable means of ensuring behavioral integrity in General Purpose AI (GPAI) systems. This connection between interpretability and alignment positions XAI metrics as both technical and regulatory instruments that help prevent *alignment faking*, a growing concern among oversight bodies. We propose a **Governance-by-Metrics** paradigm that treats explainability evaluation as a central mechanism of private AI governance. Our framework introduces a hierarchical model linking transparency, tamper resistance, scalability, and legal alignment, extending evaluation from model introspection toward systemic accountability. Through conceptual synthesis and alignment with governance standards, we outline a roadmap for integrating explainability metrics into continuous AI assurance pipelines that serve both private oversight and regulatory needs.

1 Introduction

As AI systems evolve into generative and agentic architectures, the reliability of explainability metrics becomes a governance issue—determining whether systems can be audited, trusted, and lawfully deployed. AI is already embedded in daily life and high-stakes domains [1, 2], with applications spanning healthcare, finance, and law enforcement. As its impact grows, systems must be transparent and their decision-making explainable. We argue XAI metrics are no longer only diagnostic but also enforcement levers governing transparency, trust, and accountability across the lifecycle, helping identify and mitigate risks [3, 4, 5]. Yet a critical gap persists: **we lack standardized, reliable metrics to evaluate the effectiveness and trustworthiness of explanations**.

The evaluation landscape is fragmented and often subjective [6], enabling manipulation and weakening comparisons across tasks [7, 8]. To be effective in high-risk settings, metrics must reliably assess fidelity, robustness, and usability. While regulatory frameworks like the EU AI Act [9, 10] and ISO 42001 [11] provide legal baselines, effective AI governance also depends on private oversight mechanisms—technical and institutional processes within organizations that continuously monitor and verify compliance. This aligns with emerging perspectives on private AI governance, where

First EurIPS Workshop on Private AI Governance held alongside EurIPS 2025.

assurance bodies, insurers, and certification consortia use quantitative evaluation for continuous oversight. XAI metrics can anchor these mechanisms by providing verifiable signals of explainability quality for use in insurance underwriting, procurement, and certification.

This challenge extends to General Purpose AI (GPAI) systems, where interpretability-based alignment has shown measurable potential [12, 13]. These findings suggest that explainability metrics can move beyond evaluation toward behavioral steering—a crucial capability for both private assurance and regulatory compliance. As GPAI systems become more complex and autonomous, the need for governance-aligned interpretability becomes critical for ensuring these systems remain accountable and controllable.

Position: Reliable XAI metrics are essential for technical progress and both private and regulatory compliance. Advanced systems—particularly General Purpose AI (GPAI) models—demand scalable, manipulation-resistant evaluation aligned with real-world governance needs. We call for metrics that are contextually adaptive and serve as instruments for private oversight. By building robust core metrics and aligning them with both regulatory frameworks and private governance mechanisms, we can establish consistent, useful evaluation. Collaboration among researchers, industry, and regulators is key to achieving meaningful, trustworthy, and compliant explanations.

The remainder of this paper is structured as follows: Section 2 provides background on XAI methods and evaluation metrics; Section 3 identifies key challenges; Section 4 examines mechanistic interpretability advances; Sections 5-6 establish requirements and alternative views; Section 7 outlines our governance roadmap; Section 8 discusses policy integration; Section 9 concludes; and Section 10 provides an impact statement.

2 Background and Context

Explainable AI (XAI) aims to make complex ML models understandable [14], which is vital in high-impact domains like healthcare, finance, and law enforcement. Beyond accuracy, systems must provide transparent, accountable, and comprehensible explanations—especially where outcomes are consequential [15].

Interpretability, explainability, and feature attribution serve related but distinct goals [16]. Interpretability concerns how readily humans can follow a model's reasoning; explainability provides reasons for predictions (e.g., LIME [17], SHAP [18]); attribution quantifies each input's contribution to an output. Together, they advance accountability and trust.

2.1 Explainability Methods

Intrinsic explainability [19, 20] uses inherently interpretable models (e.g., decision trees, linear regression, rule-based systems). These expose decision logic directly but may trade accuracy for simplicity. Recent work questions the faithfulness of some "inherently interpretable" claims [21], and attention-based explanations have faced criticism [22, 6].

Post hoc explainability [19, 23] explains trained models without changing them. Notable methods include SHAP [18], LIME [17], Grad-CAM [24], and Integrated Gradients [25]. These methods are flexible and widely used but approximate model behavior.

These approaches face limitations: approximations can miss true behavior; explanations may be inconsistent under small input changes; and they can be manipulated adversarially [26]. Moreover, with no control over the model, faithfulness can be difficult [6].

Example-based methods compare similar cases; mechanistic interpretability reverse-engineers internals [27, 28].

2.2 The Role of Evaluation Metrics in XAI

While explainability methods have advanced considerably, their evaluation remains inconsistent, hindering reproducibility and comparability. Recent progress in quantitative analysis introduced a broad set of evaluation metrics [29] for assessing reliability and effectiveness [30]. These help researchers and practitioners assess how well explanations reflect model decision-making and meet requirements like transparency, robustness, and usability. Private governance initiatives increasingly rely on measurable indicators to inform assurance practices—analogous to how financial auditing depends on standardized accounting principles. Reliable XAI metrics could thus serve as "auditing primitives" for model interpretability and robustness. Over time, it has become clear that most XAI metrics can be grouped into six categories:

1. **Faithfulness:** Metrics measure how well explanations reflect the model's true decision-making process, ensuring accuracy and alignment with actual predictions [31, 32].
2. **Robustness:** Metrics evaluate stability under varying inputs, including adversarial attacks and perturbations, ensuring explanations maintain integrity across test conditions [33, 34].
3. **Localisation:** Metrics assess the ability to highlight relevant regions or features that most influence model decisions, particularly important for image data [35, 36].
4. **Complexity:** Metrics evaluate simplicity and comprehensibility, ensuring explanations are accessible to end-users without unnecessary complexity [37].
5. **Randomisation (Sensitivity):** Metrics examine sensitivity to input data or parameter changes, ensuring explanations don't rely on trivial variations [38, 39].
6. **Axiomatic:** Metrics evaluate inherent properties like consistency, completeness, and preservation across architectures, grounded in theoretical foundations [25, 40].

2.3 Existing Evaluation Frameworks and Benchmarking Tools for Model Explainability

Current evaluation frameworks face significant issues. Many metrics don't capture real-world model complexity, and benchmarks lack flexibility across domains. Research suggests shifting focus toward robustness, generalizability, and actionability, while accounting for model evolution.

M-4 Benchmark [41] and OpenXAI [42] address some gaps but have limitations: M4 focuses on faithfulness without robustness, OpenXAI relies on synthetic data, and Quantus [8] struggles with human judgment alignment. Other tools like Captum [43] focus on fairness and attribution but lack standardized comparison methods.

Specialized libraries like Ferret [44] and Inseq [45] offer contributions: Ferret examines post-hoc methods but is limited to text models, while Inseq targets NLP sequence generation tasks.

2.4 Distinctive Contributions and Novelty

Unlike prior calls such as M4 [41] and OpenXAI [42], which emphasize benchmark standardization, our position contributes three new governance-oriented dimensions:

1. **Tamper-Resistance:** We propose manipulation-resilient metrics that decouple evaluation hyperparameters from model-specific artifacts, establishing auditability.
2. **Regulatory Alignment:** We explicitly map metric requirements to global governance frameworks (EU AI Act, NIST AI RMF [46], ISO/IEC 42001 [11]).
3. **Cross-Modality Integration:** We extend reliability evaluation to multimodal and agentic systems, capturing decision-making dependencies across modalities and agents.
4. **Private Governance Integration:** We position XAI metrics as quantitative instruments for private oversight—enabling certification, liability assessment, and procurement-based governance of AI systems, complementing statutory regulation.

By embedding governance principles into metric design, this framework transforms evaluation from passive assessment to active oversight, offering a distinct governance-by-design pathway that serves both private and public governance needs.

With this foundation established, we next examine the critical challenges that hinder current XAI evaluation practices.

3 Challenges in XAI Metrics

Evaluating XAI methods presents several challenges that hinder reliability and adoption. These issues prevent effective, universally applicable evaluation frameworks needed by regulators, risk managers, and users. Major problems include fragmentation, subjectivity, and manipulation vulnerabilities. Heavy reliance on hyperparameters creates opportunities to tune for favorable results [7, 47].

3.1 Neglect of Modern AI Models

Current XAI evaluation metrics struggle to capture the complexity of modern AI models, particularly large language models and autoregressive systems. These models rely on intricate decision-making processes and large-scale architectures, requiring more adaptable evaluation methods.

Explainability research has mainly focused on image and tabular modalities, with recent efforts extending to NLP. However, multi-modal AI systems are becoming increasingly common, yet most XAI methods remain single-modality focused, limiting their applicability to models processing text, images, and structured data together. Among post-hoc explanation methods, only a few, such as Layer-wise Relevance Propagation [48] and DLBacktrace [49], extend to multi-modal settings, leaving a significant evaluation gap.

As multi-modal AI adoption grows, the lack of standardized evaluation frameworks hinders interpretability and trustworthiness across domains. Existing XAI methods fail to capture dependencies between modalities. For instance, vision-based methods like Grad-CAM fail to explain text contributions in vision-language models like CLIP, while SHAP and LIME overlook image-based reasoning. This limitation is particularly critical in applications such as medical AI, where decisions rely on both textual reports and diagnostic images. Without dedicated multi-modal explainability metrics, evaluating these models remains inconsistent and unreliable. Addressing this challenge requires new faithfulness metrics that measure explanation alignment across modalities, along with benchmarking datasets to establish industry-wide standards.

3.2 Fragmentation and Manipulation Risks

A key challenge is the fragmentation in XAI evaluation due to the lack of standardized frameworks. Different metrics are used across studies, making comparing results and drawing broad conclusions hard. This inconsistency slows progress and hinders the development of best practices. Without a unified system, the field lacks direction, limiting collaboration and advancements in XAI. This fragmentation mirrors the broader governance challenge—without standardized metrics, explainability cannot serve as an accountability instrument.

XAI metrics are vulnerable to intentional or unintentional manipulation, which undermines trust in XAI systems and diminishes their practical value. Examples include adjusting evaluation parameters to achieve desired outcomes, optimizing explanations to perform well on specific metrics while not accurately reflecting the model's true decision-making process, and using adversarial inputs to create explanations that seem robust but fail in real-world applications. Without tamper-resistant benchmarks, evaluation risks devolving into *explanation theater*, weakening both private and public trust. As recent studies show, the sensitivity of XAI evaluations to hyperparameters—whether tied to model architecture, attribution baselines, or metric parameters—creates pathways for manipulation. For example, baseline selection in Integrated Gradients or perturbation order in faithfulness metrics can drastically shift rankings [50, 25]. This interdependence highlights the need for tamper-resistant evaluation frameworks capable of maintaining metric stability under parameter variation.

Recent research has increasingly focused on the role of hyperparameters in XAI evaluations and how they can introduce confounding effects [8]. Studies have explored how sensitive attribution methods are to explanation hyperparameters like random seed or sample size [51], how baseline choices impact explanation outcomes [50, 25], and how model changes (optimizer, activation, learning rate, data splits) influence explanations [52]. Research has also investigated how normalization, randomization order, and similarity measures affect evaluation outcomes [53, 54].

To address these challenges and establish reliable governance-aligned metrics, we must first examine the technical foundation provided by recent advances in mechanistic interpretability. These developments demonstrate how interpretability can move beyond passive explanation toward active alignment and governance.

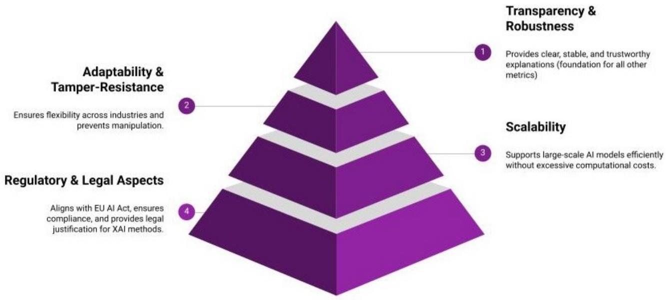
Transparency & Robustness
1
Provides clear, stable, and trustworthy explanations (foundation for all other metrics)
Adaptability & Tamper-Resistance
2
Ensures flexibility across industries and prevents manipulation.
Scalability
3
Supports large-scale AI models efficiently without excessive computational costs.
Regulatory & Legal Aspects
4
Aligns with EU AI Act, ensures compliance, and provides legal justification for XAI methods.

Figure 1: Governance Hierarchy for XAI Evaluation. This pyramid illustrates foundational Transparency & Robustness, building up through Adaptability & Tamper-Resistance and Scalability, to Regulatory & Legal Aspects.

4 Mechanistic Interpretability and Alignment for Governance

Recent interpretability efforts such as TransformerLens [55], CircuitsVis [56], and Scaling Monosemanticity [57] reveal neuron-level insights into large models, yet they lack measurable governance alignment. Recent advances [12, 13] demonstrate that interpretability can *directly guide model alignment*, showing that pruning and attribution-aware feedback can enhance both efficiency and ethical behavior. Such interpretability-guided alignment not only enhances model integrity but also provides measurable assurance to private auditors verifying alignment claims. This evidence supports interpretability as a concrete mechanism for aligning General Purpose AI (GPAI) systems under regulatory scrutiny.

Regulators have increasingly cited "alignment faking" — models simulating ethical behavior without genuine internal compliance — as a critical governance risk [58]. Embedding interpretability-guided evaluation within assurance pipelines provides an empirical safeguard against such deception, bridging technical transparency with legal accountability. Our framework extends evaluation into agentic and generative systems, establishing metrics to assess transparency, autonomy, and alignment drift for private auditors and certification bodies. As AI systems become more complex and autonomous, the need for governance-aligned interpretability becomes critical for ensuring these systems remain accountable and controllable. Mechanistic interpretability explains what a model does; governance-oriented evaluation ensures that what it does remains accountable to both private oversight mechanisms and regulatory frameworks.

Building on these mechanistic interpretability advances, we now establish the key requirements that reliable XAI metrics must satisfy to serve as effective governance instruments for both private and public oversight.

5 Key Requirements for Reliable Metrics

To overcome these challenges, reliable XAI metrics must be developed. These metrics should establish standard benchmarks for explainability, helping to compare, quantify, and qualify evaluation results. Standardization will clarify regulatory requirements, minimizing bias in user-preferred choices [59, 60]. As depicted in Figure 1, reliable metrics must meet the following criteria:

Transparency & Robustness : XAI evaluation metrics must provide clear, consistent insights into how well an explanation aligns with the model's decision-making process. Transparent metrics ensure stakeholder trust. Metrics should assess the stability of explanations under various conditions, including adversarial inputs and data changes [61]. A robust metric ensures that explanations hold up under real-world variations, offering consistency and reliability across different environments. For example, a private auditor evaluating a credit scoring model would require transparent metrics showing not only which features influenced decisions but also how stable those influences remain across different applicant populations and temporal periods.

Adaptability & Tamper-Resistance : XAI frameworks should be flexible enough to cater to diverse domains, such as healthcare or finance, where priorities like interpretability or compliance vary

[61]. Adaptable metrics ensure their applicability across different sectors, addressing the specific challenges of each. Moreover, these metrics must incorporate safeguards such as adversarial testing and regular validation to prevent manipulation, ensuring that explanations are not artificially adjusted to meet predetermined standards. A healthcare certification body, for instance, must ensure that explainability metrics for diagnostic AI adapt to clinical workflows while remaining resistant to manipulation through hyperparameter tuning—a balance essential for both usability and integrity.

Scalability : Scalable explainability metrics are critical for evaluating modern AI systems, particularly large-scale models such as LLMs. Existing explainability metrics are often computationally expensive, limiting their feasibility for large-scale deployments. As AI models grow in complexity, it is imperative to develop scalable methods that can provide meaningful explanations without excessive computational overhead [62]. Certification platforms can batch-evaluate hundreds of models using scalable explainability benchmarks, enabling efficient assessment across diverse model types and sizes, from small specialized models to large foundation models, without prohibitive computational costs.

Regulatory & Legal Aspects : While the EU AI Act formalizes transparency and accountability obligations, parallel governance frameworks such as NIST's AI Risk Management Framework (2023) [46], ISO/IEC 42001 (2024) [11], and Singapore's Model AI Governance Framework [63] embody equivalent principles of auditability and traceability. XAI metrics must align with regulatory frameworks such as the EU AI Act, which mandates transparency and interpretability for high-risk AI systems [64, 10]. These metrics must help organizations demonstrate compliance by providing clear, traceable, and auditable explanations. Cross-jurisdictional audit value is enhanced when standardized metrics enable consistent evaluation across different regulatory regimes, facilitating international AI governance cooperation. From a legal perspective, XAI metrics must ensure that AI decisions are explainable and justifiable in court [65], particularly in healthcare and finance sectors where AI decisions have significant consequences [66, 67].

Together, these criteria define a governance hierarchy that connects technical transparency to regulatory verifiability. These requirements inform our governance-by-design approach, which we detail in the following section.

6 Alternative Views

Needs differ by domain (e.g., clinical vs. financial), so one-size-fits-all metrics can hinder effectiveness [68]. Expert qualitative evaluation complements quantitative metrics to capture context and nuance [69, 70]. A hybrid approach combines core benchmarks (e.g., fidelity, robustness) with domain-specific metrics and human input [71].

Industry stakeholders emphasize practical implementation challenges of private governance mechanisms. Technology companies advocate for flexible, market-driven approaches that allow innovation while maintaining accountability [72]. Insurance providers and certification bodies highlight the need for standardized metrics that enable risk assessment and liability determination across diverse AI applications [73, 74]. This diversity of perspectives underscores the importance of adaptable evaluation frameworks that can accommodate sector-specific requirements while maintaining core governance principles.

Governance approaches vary significantly across jurisdictions, reflecting different cultural and legal traditions. While the EU emphasizes prescriptive regulatory frameworks, other regions favor principles-based approaches that rely more heavily on private governance mechanisms [63]. This variation creates both challenges and opportunities: challenges in achieving cross-border interoperability, but opportunities for regulatory experimentation and learning. Our governance-by-metrics approach provides a common technical foundation that can adapt to diverse regulatory contexts while maintaining consistent evaluation standards.

Our framework acknowledges these diverse perspectives by positioning XAI metrics as flexible governance instruments rather than rigid compliance checklists. The hierarchical model (Figure 1) accommodates domain-specific adaptations while ensuring core requirements—transparency, robustness, tamper-resistance—remain non-negotiable. Nonetheless, over-standardization may stifle methodological innovation if not balanced with domain flexibility. This balance between standardization and flexibility enables both private governance innovation and regulatory compliance.

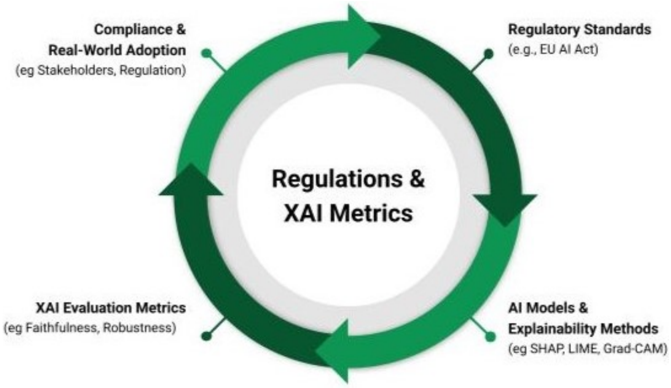

Figure 2: The Iterative Feedback Loop in XAI Regulation and Governance Compliance. Regulatory standards influence XAI evaluation metrics and AI models; subsequent real-world adoption and challenges then refine these standards and metrics to align with evolving AI capabilities.

Having established these requirements, we now present a comprehensive roadmap for implementing governance-by-metrics in practice.

7 Governance-by-Design: A Three-Phase Roadmap

To address the current gaps in XAI evaluation and ensure the development of reliable, transparent, and compliant AI systems, we propose a three-phase governance roadmap that embeds accountability into the evaluation process:

1. **Phase I: Metric Integrity** Develop transparent, reproducible faithfulness metrics and a public integrity registry, covering fidelity, robustness, clarity, and comprehensibility.
2. **Phase II: Private Assurance & Certification** Embed metrics into audits, insurance workflows, and certification processes, enabling private governance actors to assess and verify AI system trustworthiness.
3. **Phase III: Regulatory Interoperability** Align private metrics with legal frameworks to ensure traceable compliance across jurisdictions, scalable to advanced models including LLMs, while maintaining compatibility with private governance mechanisms.

Sectors require tailored emphases (e.g., clinical confidence vs. financial compliance and auditability). Explainability methods can be computationally heavy, limiting use on large models. Evaluation must balance efficiency with fidelity to keep interpretability feasible at scale [75]. Collaboration among academia, industry, and regulators is vital to align metrics with real-world needs, especially in high-risk areas [76].

For instance, assurance bodies could use standardized explainability metrics to certify model transparency, enabling interoperability across markets while maintaining competitive flexibility. This phased roadmap is among the first to operationalize governance principles through measurable XAI metrics, enabling continuous accountability for agentic and generative AI systems.

This roadmap's success depends on effective integration with broader policy frameworks and private governance mechanisms. We now examine how XAI metrics can be embedded within existing and emerging governance structures.

8 Policy Integration: Private and Public Governance Mechanisms

XAI metrics serve as critical instruments for both private and public governance, enabling accountability, trust, and compliance across diverse institutional contexts.

8.1 Private Governance Mechanisms

Third-Party Auditing and Certification: Metrics quantify explanation quality for auditors and evaluators [77, 78], enabling independent verification by private auditors and certification bodies. Standardized evaluation frameworks allow these actors to assess model trustworthiness across different domains and use cases. For instance, an AI certification body might use standardized explainability metrics to assess whether a healthcare diagnostic model meets transparency requirements before hospital procurement. The certification process would evaluate faithfulness scores, robustness under distribution shift, and consistency across patient demographics—providing hospitals with verifiable assurance of model trustworthiness.

Insurance and Liability Assessment: Private governance leverages metrics as market instruments—through certification schemes, transparency indices, and risk-adjusted insurance models. These mechanisms enable self-regulation that complements formal oversight, promoting responsible AI through economic and reputational incentives. Consider an AI liability insurance policy: insurers could use explainability metrics to assess risk exposure, offering lower premiums for models demonstrating high faithfulness and robustness scores. This creates market incentives for developers to prioritize explainability, complementing regulatory mandates with economic drivers for responsible AI development.

Procurement-Based Governance: Objective criteria help organizations demonstrate compliance and foster trustworthy deployment [9], enabling procurement-based governance where purchasing decisions incorporate explainability requirements.

8.2 Regulatory Integration

Regulatory Compliance: XAI metrics align with legal frameworks (EU AI Act, NIST AI RMF, ISO/IEC 42001) to create auditable templates that satisfy both technical and legal standards.

Auditability and Traceability: Metrics provide clear, consistent evaluation that improves user confidence, especially in high-stakes domains [79], while ensuring legal traceability and accountability.

Building Trust: Prioritizing fairness, robustness, and transparency aligns XAI with societal values [80, 81], fostering public confidence in AI systems.

8.3 GPAI Oversight and Cross-Jurisdictional Governance

As governance frameworks evolve, oversight of General Purpose AI (GPAI) models will depend increasingly on measurable interpretability and alignment metrics. Private governance actors—including insurers, AI assurance labs, and model marketplaces—play crucial roles in GPAI evaluation, suggesting a need for common explainability metrics that transcend legal boundaries while remaining auditable. For example, a foundation model marketplace might require explainability scores as part of model listings, enabling downstream users to assess transparency before integration. An AI assurance lab could provide independent verification of these scores, creating a trust infrastructure for GPAI deployment that transcends individual regulatory jurisdictions. Integrating interpretability-based alignment into private governance infrastructures provides a safeguard against alignment faking—ensuring that models not only appear compliant but demonstrably behave in alignment with ethical and legal standards, providing scalable, jurisdiction-agnostic oversight that complements formal regulation.

8.4 Broader Societal Implications

Advancing Ethical AI: Standardized benchmarks enable comparison, best practices, and faster progress. Regulation and explainability co-evolve (Figure 2), creating a virtuous cycle where technical advances inform governance and governance requirements drive technical innovation.

9 Conclusion

Reliable explainability is the foundation of both public and private algorithmic governance. By integrating XAI metrics with global frameworks—NIST AI RMF [46], ISO/IEC 42001 [11], OECD AI Principles [82], and Singapore's Model AI Governance Framework [63]—this work defines a path toward measurable, auditable accountability. The Governance-by-Metrics paradigm thus extends explainability research into the domain of *AI governance engineering*—embedding transparency, accountability, and legal verifiability into the very fabric of model evaluation.

Interpretability-guided alignment demonstrates that reliable explanations serve not only as diagnostic tools but as behavioral governance instruments for GPAI systems, reinforcing the central role of XAI metrics in both private assurance and regulatory compliance. Despite progress in explainability methods, evaluation remains fragmented, subjective, and prone to manipulation. Key challenges include lack of standardization, manipulation risks, limited multi-modal support, and regulatory misalignment. Without rigorous evaluation, explainability risks becoming a mere regulatory formality. Advancing beyond theoretical explainability toward standardized, governance-integrated evaluation frameworks will ensure AI systems remain both accountable and compliant across jurisdictions.

10 Impact Statement

Reliable XAI metrics enable cost-efficient auditing, certification, and insurance underwriting for AI systems. By embedding interpretability within private governance infrastructures, these metrics ensure scalable, jurisdiction-agnostic oversight of General Purpose AI. Integrating alignment and auditability within evaluation transforms explainability into an operational mechanism for responsible AI deployment.

References

[1] OpenAI. OpenAI ChatGPT. https://openai.com/index/chatgpt/. [Accessed 12-11-2024].

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

[3] Dario Amodei, Christopher Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dandelion Mané. Concrete problems in ai safety. *arXiv preprint*, abs/1606.06565, 2016. URL https://api.semanticscholar.org/CorpusID:10242377.

[4] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

[5] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint*, abs/1811.07871, 2018. URL https://api.semanticscholar.org/CorpusID:53745764.

[6] Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. Interpretability needs a new paradigm, 2024. URL https://arxiv.org/abs/2405.05386.

[7] Kristoffer Wickstrøm, Marina Marie-Claire Höhne, and Anna Hedström. From flexibility to manipulation: The slippery slope of xai evaluation, 2024. URL https://arxiv.org/abs/2412.05592.

[8] Anna Hedström, Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M. C. Höhne. Quantus: An explainable

ai toolkit for responsible evaluation of neural network explanations and beyond, 2023. URL https://arxiv.org/abs/2202.06861.

[9] Regulation - EU - 2024/1689 - EN - EUR-Lex — eur-lex.europa.eu. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689. [Accessed 31-01-2025].

[10] Luca Nannini, Jose Maria Alonso-Moral, Alejandro Catalá, Manuel Lama, and Senén Barro. Operationalizing explainable artificial intelligence in the european union regulatory ecosystem. *IEEE Intelligent Systems*, 39(4):37–48, 2024. doi: 10.1109/MIS.2024.3383155.

[11] International Organization for Standardization. Iso/iec 42001:2023 information technology — artificial intelligence — management system, 2023. ISO/IEC Standard.

[12] A. Malik et al. Interpretability-aware pruning for efficient medical ai, 2025. URL https://arxiv.org/abs/2509.08592. Under review.

[13] Aadit Sengupta, Pratinav Seth, and Vinay Kumar Sankarapu. Interpretability-guided alignment for general purpose ai, 2025.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019. URL https://arxiv.org/abs/1910.10045.

[15] Yan Jia, John McDermid, Tom Lawton, and Ibrahim Habli. The role of explainability in assuring safety of machine learning in healthcare, 2022. URL https://arxiv.org/abs/2109.00520.

[16] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*, 2 2017. URL http://arxiv.org/abs/1702.08608.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. URL https://api.semanticscholar.org/CorpusID:13029170.

[18] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:21889700.

[19] Zachary C Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10): 36–43, 9 2018. ISSN 0001-0782. doi: 10.1145/3233231. URL https://dl.acm.org/doi/10.1145/3233231.

[20] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL http://www.nature.com/articles/s42256-019-0048-x.

[21] Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Stroudsburg, PA, USA, 4 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://www.aclweb.org/anthology/2020.acl-main.386.

[22] Sofia Serrano and Noah A Smith. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Stroudsburg, PA, USA, 6 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://www.aclweb.org/anthology/P19-1282.

[23] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Computing Surveys*, 55(8):1–42, 8 2022. ISSN 0360-0300. doi: 10.1145/3546577. URL https://dl.acm.org/doi/10.1145/3546577.

[24] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336 – 359, 2016. URL `https://api.semanticscholar.org/CorpusID:15019293`.

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017. URL `https://api.semanticscholar.org/CorpusID:16747630`.

[26] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods, 2020. URL `https://arxiv.org/abs/1911.02508`.

[27] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

[28] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint*, abs/2301.05217, 2023. URL `https://api.semanticscholar.org/CorpusID:255749430`.

[29] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=MU2495w47rz`.

[30] Md Kadir, Amir Mosavi, and Daniel Sonntag. Evaluation metrics for xai: A review, taxonomy, and practical applications. pages 000111–000124, 07 2023. doi: 10.1109/INES59282.2023.10297629.

[31] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *International Joint Conference on Artificial Intelligence*, 2020. URL `https://api.semanticscholar.org/CorpusID:218486810`.

[32] David Alvarez-Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. *arXiv preprint*, abs/1806.07538, 2018. URL `https://api.semanticscholar.org/CorpusID:49324194`.

[33] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations, 2019. URL `https://arxiv.org/abs/1901.09392`.

[34] Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations, 2022. URL `https://arxiv.org/abs/2203.06877`.

[35] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with lrp, 2020. URL `https://arxiv.org/abs/1910.09840`.

[36] Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, May 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.11.008. URL `http://dx.doi.org/10.1016/j.inffus.2021.11.008`.

[37] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Somesh Jha, and Xi Wu. Concise explanations of neural networks using adversarial training, 2020. URL `https://arxiv.org/abs/1810.06583`.

[38] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2020. URL `https://arxiv.org/abs/1810.03292`.

[39] Anna Hedström, Leander Weber, Sebastian Lapuschkin, and Marina M.-C. Höhne. Sanity checks revisited: An exploration to repair the model parameter randomisation test. *arXiv preprint*, abs/2401.06465, 2024. URL https://api.semanticscholar.org/CorpusID:266977162.

[40] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods, 2017. URL https://arxiv.org/abs/1711.00867.

[41] Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. $\mathcal{M}^4$: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=6zcfrSz98y.

[42] Chirag Agarwal, Dan Ley, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations, 2024. URL https://arxiv.org/abs/2206.11104.

[43] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. URL https://arxiv.org/abs/2009.07896.

[44] Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, May 2023.

[45] Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-demo.40. URL http://dx.doi.org/10.18653/v1/2023.acl-demo.40.

[46] National Institute of Standards and Technology. Ai risk management framework (ai rmf 1.0). Technical report, U.S. Department of Commerce, 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[47] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2019. URL https://arxiv.org/abs/1812.04608.

[48] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: Attention-aware layer-wise relevance propagation for transformers, 2024. URL https://arxiv.org/abs/2402.05602.

[49] Vinay Kumar Sankarapu, Chintan Chitroda, Yashwardhan Rathore, Neeraj Kumar Singh, and Pratinav Seth. Dlbacktrace: A model agnostic explainability for any deep learning models, 2024. URL https://arxiv.org/abs/2411.12643.

[50] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. URL https://distill.pub/2020/attribution-baselines/.

[51] Naman Bansal, Chirag Agarwal, and Anh Nguyen. SAM: the sensitivity of attribution methods to hyperparameters. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 11–21. Computer Vision Foundation / IEEE, 2020.

[52] Amir-Hossein Karimi, Krikamol Muandet, Simon Kornblith, Bernhard Schölkopf, and Been Kim. On the relationship between explanation and prediction: A causal view. In *XAI in Action: Past, Present, and Future Applications*, 2023. URL https://openreview.net/forum?id=ag1CpSUjPS.

[53] Anna Hedström, Leander Weber, Sebastian Lapuschkin, and Marina Höhne. A fresh look at sanity checks for saliency maps. In *Explainable Artificial Intelligence*, pages 403–420, Cham, 2024. Springer Nature Switzerland.

[54] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods, 2022. URL https://arxiv.org/abs/2202.00449.

[55] Neel Nanda, Lawrence Chan, Tom Lieberum, Jesse Smith, and Jacob Steinhardt. Transformerlens: A library for mechanistic interpretability of transformer language models, 2022. URL https://github.com/neelnanda-io/TransformerLens.

[56] Callum Cooney et al. Circuitsvis: A library for visualizing neural network circuits, 2022. URL https://github.com/alan-cooney/CircuitsVis.

[57] Tom Templeton et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024.

[58] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martinez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback. *Ethics and Information Technology*, 27, 2025. URL https://api.semanticscholar.org/CorpusID:279163270.

[59] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, July 2023. ISSN 1557-7341. doi: 10.1145/3583558. URL http://dx.doi.org/10.1145/3583558.

[60] Aleksandra Pawlicka, Marek Pawlicki, Rafał Kozik, Wiktor Kurek, and Michał Choraś. *How Explainable Is Explainability? Towards Better Metrics for Explainable AI*, pages 685–695. 01 2024. ISBN 978-3-031-44720-4. doi: 10.1007/978-3-031-44721-1_52.

[61] Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 45–50, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

[62] Jean Dessain, Nora Bentaleb, and Fabien Vinas. Cost of explainability in ai: An example with credit scoring models. pages 498–516, 10 2023. ISBN 978-3-031-44063-2. doi: 10.1007/978-3-031-44064-9_26.

[63] Personal Data Protection Commission Singapore. Model ai governance framework. Technical report, Government of Singapore, 2019. URL https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework.

[64] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. Metrics, explainability and the european ai act proposal. *J*, 5(1):126–138, 2022. ISSN 2571-8800. doi: 10.3390/j5010010. URL https://www.mdpi.com/2571-8800/5/1/10.

[65] Fabian Walke, Lars Bennek, and Till Winkler. Artificial intelligence explainability requirements of the ai act and metrics for measuring compliance. 09 2023.

[66] Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco F. Huber, and Christian Horz. How should ai decisions be explained? requirements for explanations from the perspective of european law. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:438–450, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31648. URL http://dx.doi.org/10.1609/aies.v7i1.31648.

[67] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. Impact of legal requirements on explainability in machine learning, 2020. URL https://arxiv.org/abs/2007.05479.

[68] Essi Pietilä and Pedro Moreno. When an explanation is not enough: An overview of evaluation metrics of explainable ai systems in the healthcare domain. pages 573–584, 01 2024. ISBN 978-3-031-49061-3. doi: 10.1007/978-3-031-49062-0_60.

[69] Upol Ehsan and Mark Riedl. On design and evaluation of human-centered explainable ai systems [position paper]. 04 2019.

[70] Julien Colin, Thomas Fel, Remi Cadene, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods, 2023. URL `https://arxiv.org/abs/2112.04417`.

[71] Shuai Ma. Towards human-centered design of explainable artificial intelligence (xai): A survey of empirical studies, 2024. URL `https://arxiv.org/abs/2410.21183`.

[72] Dean W. Ball. A framework for the private governance of frontier artificial intelligence, 2025. URL `https://arxiv.org/abs/2504.11501`.

[73] Gabriel Weil, Matteo Pistillo, Suzanne Van Arsdale, Junichi Ikegami, Kensuke Onuma, Megumi Okawa, and Michael A. Osborne. Insuring emerging risks from ai, 2024. URL `https://www.oxfordmartin.ox.ac.uk/publications/insuring-emerging-risks-from-ai`.

[74] Anat Lior. Innovating liability: The virtuous cycle of torts, technology and liability insurance. *Yale Journal of Law and Technology*, 25:448, 2023. URL `https://yjolt.org/innovating-liability-virtuous-cycle-torts-technology-and-liability-insurance`.

[75] Claire Jean-Quartier, Katharina Bein, Lukas Hejny, Edith Hofer, Andreas Holzinger, and Fleur Jeanquartier. The cost of understanding - xai algorithms towards sustainable ml in the view of computational cost. *Comput.*, 11:92, 2023. URL `https://api.semanticscholar.org/CorpusID:258542839`.

[76] Beyond silos: Why AI Regulation calls for an Interdisciplinary Approach | Feature from King&apos;s College London — kcl.ac.uk. `https://www.kcl.ac.uk/beyond-silos-why-ai-regulation-calls-for-an-interdisciplinary-approach`. [Accessed 31-01-2025].

[77] Yueqi Li and Sanjay Goel. Making it possible for the auditing of AI: A systematic review of AI audits and AI auditability. *Inf. Syst. Front.*, July 2024.

[78] Louise McCormack and Malika Bendechache. A comprehensive survey and classification of evaluation criteria for trustworthy artificial intelligence. *AI Ethics*, October 2024.

[79] Francesco Sovrano, Fabio Vitali, and Monica Palmirani. *Making Things Explainable vs Explaining: Requirements and Challenges Under the GDPR*, page 169–182. Springer International Publishing, 2021. ISBN 9783030898113. doi: 10.1007/978-3-030-89811-3_12. URL `http://dx.doi.org/10.1007/978-3-030-89811-3_12`.

[80] Luca Nannini, Marta Marchiori Manerba, and Isacco Beretta. Mapping the landscape of ethical considerations in explainable AI research. *Ethics Inf. Technol.*, 26(3), September 2024.

[81] Mohammad Amir Khusru Akhtar, Mohit Kumar, and Anand Nayyar. Transparency and accountability in explainable AI: Best practices. In *Studies in Systems, Decision and Control*, Studies in systems, decision and control, pages 127–164. Springer Nature Switzerland, Cham, 2024.

[82] Organisation for Economic Co-operation and Development. Oecd principles on artificial intelligence, 2019. URL `https://www.oecd.org/digital/artificial-intelligence/`.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the central position—that reliable and standardized metrics are essential for evaluating XAI—and this is consistently argued throughout the paper without introducing unsubstantiated claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 3 (*Challenges in XAI Metrics*) discusses technical and practical limitations, including scalability issues, fragmentation, subjectivity, and manipulation vulnerabilities.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification:This is a conceptual position paper, not a theoretical paper; no new theorems, proofs, or formal assumptions are introduced.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include new experiments or empirical benchmarks; it synthesizes and analyzes existing work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: No new datasets or code are introduced; all discussed methods and toolkits are cited from prior public work.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not present new experiments or training runs; it discusses existing techniques at a conceptual and methodological level.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: No experiments are performed, hence no statistical testing is applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No experiments were conducted in this work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is conceptual, does not involve human subjects, and explicitly addresses ethical concerns such as risks of misuse, explanation theater, and compliance (Sections 5 and 7).

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Impact Statement (Section 10) outlines both positive impacts (auditability, compliance, trust) and negative risks (bias reinforcement, misuse, adversarial exploitation).

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No new datasets or models are released; the work is conceptual.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All prior methods, toolkits, and datasets are properly cited with their original sources. Referenced libraries such as LIME, SHAP, Captum, Quantus, and DLBacktrace are open-source and distributed under permissive licenses (e.g., MIT, Apache 2.0); their terms of use are respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets; it synthesizes existing literature.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:No crowdsourcing or human-subject research was conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human participants or personal data.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Large language models were not used as a core methodological component; they are only discussed as subjects of interpretability research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.