Interpretability as Alignment: Making Internal Understanding a Design Principle

Aadit Sengupta[*†‡]
Lexsi Labs
Mumbai, India

Pratinav Seth[*§]
Lexsi Labs
Mumbai, India

Vinay K. Sankarapu
Lexsi Labs
London, United Kingdom

Abstract

Frontier AI systems require governance mechanisms that can verify internal alignment, not just behavioral compliance. Private governance mechanisms—audits, certification, insurance, and procurement are emerging to complement public regulation, but they require technical substrates that generate verifiable causal evidence about model behavior. This paper argues that mechanistic interpretability provides this substrate. We frame interpretability not as post-hoc explanation but as a design constraint embedding auditability, provenance, and bounded transparency within model architectures. Integrating causal abstraction theory and empirical benchmarks such as MIB and LoBOX, we outline how interpretability-first models can underpin private assurance pipelines and role-calibrated transparency frameworks. This reframing situates interpretability as infrastructure for private AI governance—bridging the gap between technical reliability and institutional accountability.

1 Introduction

AI systems, particularly large language models, are increasingly deployed in high-stakes settings—healthcare, education, law, and employment. These models generate fluent outputs, but their internal workings remain opaque, making it difficult to know whether their decisions reflect sound reasoning or misaligned goals. This concern has put AI alignment at the center of technical research and public discussion [1, 2, 3, 4].

Interpretability has emerged as a key strategy for alignment. If we can understand how a model makes decisions, we can better assess whether it's behaving safely. Some work focuses on post-hoc explanations like LIME or SHAP [5, 6], while mechanistic interpretability attempts to look inside model architecture—identifying which neurons, attention heads, or circuits contribute to specific behaviors [7, 8, 9]. However, post-hoc explanations are often inconsistent or manipulable [10, 11], while mechanistic work is labor-intensive and doesn't scale to frontier models. Many interpretability methods tell us stories about what the model might be doing, without strong evidence that those stories are true in a causal sense [12].

Regulatory frameworks and behavioral alignment. The EU AI Act mandates transparency for high-risk applications, particularly for General Purpose AI (GPAI) systems. Industry has largely responded with behavioral alignment techniques such as RLHF—methods that improve outputs but leave internal logic untouched [13]. Our position reframes interpretability as infrastructure for governance: embedding accountability and auditability into model design rather than applying

[*] Co-First Authorship
[†] work done during internship at Lexsi Labs
[‡] Department of Computer Science, University of Michigan Ann Arbor, Michigan.
[§] corresponding author – pratinav.seth@lexsi.ai

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

Criterion	Post-Hoc	Mechanistic
Focus	Explains outputs after training	Explains internal components/processes
Examples	LIME, SHAP, Grad-CAM	Circuits, Activation Patching, Tracing
Nature	Correlational, approximate	Causal, structurally grounded
Scalability	Easy to scale, low overhead	Resource-intensive, less scalable
Reliability	Risk of misleading narratives	Closer to true model computation

Table 1: Comparison between post-hoc and mechanistic interpretability approaches.

them post hoc. This represents a fundamental shift from post-hoc transparency to pre-embedded accountability mechanisms, distinguishing our approach from prior interpretability frameworks that focus primarily on diagnostic capabilities rather than design constraints.

Beyond public regulation, private governance mechanisms—including third-party audits, compliance certification, risk insurance, and procurement standards—are emerging as complementary accountability structures [14, 15]. These mechanisms rely on technical substrates that can generate verifiable causal evidence about model behavior. Mechanistic interpretability offers such a substrate by enabling reproducible inspection and provenance tracking at the circuit level. This positions interpretability-first design as a foundation for private governance infrastructures that bridge technical reliability and institutional accountability.

This paper argues that internal transparency is not optional but a basic requirement for building aligned systems. We examine the limits of current methods and consider tools, benchmarks, and collaborations that might help interpretability become more robust and reliable. Without solid foundations for understanding how models think, alignment risks becoming a surface-level fix for a deeper problem.

Contributions. This paper introduces three primary contributions: (1) We introduce a conceptual bridge linking mechanistic interpretability with private governance mechanisms—framing causal interpretability as the evidentiary layer for audits, certification, and insurance. (2) We specify a governance-aware technical blueprint—interpretability-first architectures that embed audit hooks, provenance tracking, and bounded transparency. (3) We connect emerging benchmarks (MIB, LoBOX) with private oversight workflows and regulatory compliance frameworks, providing an implementation roadmap for interpretability-as-governance infrastructure.

2 Introduction to Model Interpretability

Interpretability refers to how well humans can understand a model's internal behavior—how inputs are processed, decisions are formed, and outputs are produced [16]. Explainability describes human-readable justifications for outputs, while transparency concerns access to architecture, training data, or parameters [17, 18, 19]. Alignment asks whether models behave in line with human goals and values [20].

We distinguish between intrinsic interpretability (transparent by design) and post-hoc interpretability (explaining black-box models after training) [21, 22]. Table 1 summarizes the key differences between these approaches. Post-hoc methods like LIME, SHAP, and Integrated Gradients dominate practice but are frequently misleading and manipulable [23, 11]. For example, SHAP can be gamed to attribute importance to benign features while hiding decisions based on sensitive attributes [11]. Newer approaches like DL-Backtrace [24] offer deterministic tracing without baseline selection, but the fundamental challenge remains: distinguishing genuine mechanistic insight from plausible storytelling.

Causal abstraction and representation decomposition. Following Geiger et al. [25], we adopt the causal abstraction perspective, formalizing mechanistic interpretability as discovering structural homomorphisms between model components and human-interpretable causal variables. This theoretical foundation builds on Pearl's causal hierarchy [26], enabling intervention-based testing to establish relationships between high-level interpretations and low-level mechanisms. Sparse Autoencoders (SAEs) [27] aim to decompose entangled representations into interpretable components, but feature consistency across training runs and architectures remains challenging [28]. The MIB benchmark

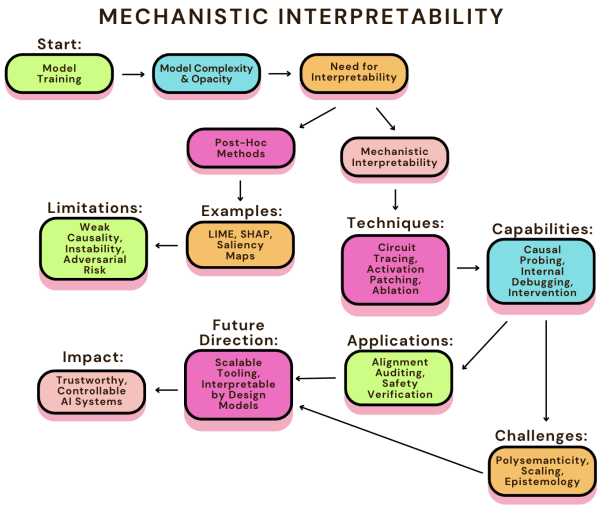

Figure 1: A high-level conceptual map of mechanistic interpretability. It contrasts post-hoc approaches with mechanistic techniques, and illustrates core techniques, applications, limitations, and future directions.

[29] provides empirical infrastructure for evaluating decomposition methods, showing that attribution approaches often outperform SAE features in circuit localization tasks.

Having established the theoretical foundations of interpretability, we now examine mechanistic interpretability as the structural basis for alignment.

3 Mechanistic Interpretability — Paradigm of Model Alignment

Mechanistic interpretability (MI) seeks to identify specific components—neurons, attention heads, or circuits—that causally contribute to model outputs [7, 8]. Recent progress in transformers reveals interpretable substructures: attention heads performing token copying, syntactic tracking, or positional induction [30, 9], and modular circuits executing string comparison and arithmetic [8]. These findings suggest that within high-parameter networks, small functional units may correspond to meaningful, testable computations. This opens the door to detecting internal failures—reward hacking or deceptive reasoning—that behavioral methods may overlook [1]. MI employs activation patching and causal tracing for controlled interventions, providing empirical insight into internal mechanisms [7].

However, MI faces significant challenges. **Polysemanticity**—individual neurons encoding multiple unrelated features—complicates semantic interpretation and becomes more pronounced at scale [8, 7]. This does not always imply superposition; polysemanticity may arise from non-linear mixtures or compositional features [31]. Recent work by Meloux et al. [32] and Sutter et al. [33] questions the identifiability of mechanistic interpretations, suggesting that multiple valid explanations may exist for the same model behavior. SAEs show promise for disentangling features but face consistency challenges across training runs and architectures [27, 28]. **Scalability** remains a bottleneck: MI requires extensive computational resources and expert labor, and tools like activation patching don't yet scale to frontier models [34, 35]. **Epistemic concerns** include confirmation bias in human pattern recognition and "explanation theater"—compelling narratives that fail under scrutiny [36, 37, 23]. This poses particular risks for governance applications, where explanation theater could undermine audit compliance and verification protocols, leading to false confidence in model safety.

Nonetheless, MI offers unique capabilities for alignment. Figure 1 provides a conceptual overview of mechanistic interpretability approaches and their applications. Behavioral methods like RLHF focus on outputs without addressing internal reasoning, potentially leaving unsafe or deceptive processes intact [3]. MI provides tools for interrogating and modifying internal processes, enabling alignment at the reasoning level rather than just performance. This positions MI as essential for building auditable, verifiable AI systems [38, 20]. By exposing causal pathways and enabling targeted interventions, MI supports governance frameworks like the EU AI Act while respecting bounded-opacity principles

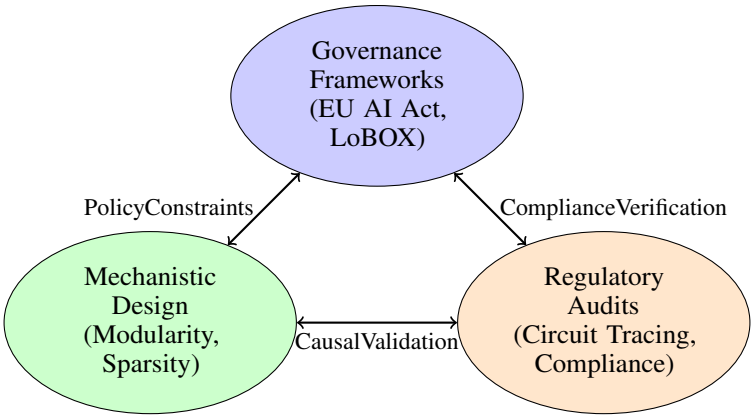

Figure 2: Interpretability-Alignment-Governance Triangle: Bidirectional relationships between mechanistic design principles, regulatory audit capabilities, and governance frameworks. Each component constrains and enables the others, creating a feedback loop for interpretability-first AI development.

[39]. Future progress depends on scalable toolchains, robust benchmarks, and hybrid approaches combining mechanistic insights with behavioral fine-tuning [40, 37].

The capabilities discussed above point toward a broader vision: interpretability as a design principle for alignment rather than a post-hoc diagnostic tool.

4 Interpretability as a Design Principle for Private Governance

Interpretability should be viewed not only as a mechanism for aligning AI systems with human intent, but as the technical substrate enabling private governance mechanisms that complement public regulation. By exposing and intervening on internal representations, interpretability enables verifiable causal evidence for third-party audits, certification bodies, and risk assessment frameworks [41, 21, 3].

Architectural desiderata. Interpretability-first design requires modularity for component-level inspection, sparsity to reduce polysemanticity, controlled polysemanticity bounding features per unit, audit hooks for immutable state records, intervention-friendliness supporting surgical edits, and provenance tracking maintaining representation lineage. These constraints represent constraints during model training and architecture design, not post-hoc adaptations. For example, a model with traceable modular circuits could allow a regulator to confirm that decision rules comply with anti-discrimination norms without retraining, demonstrating how interpretability-guided alignment serves as proof-of-concept for General Purpose AI (GPAI) governance requirements. Recent work on interpretability-aware pruning [42] shows how architectural constraints can be embedded during training rather than applied post-hoc. These constraints trade off against raw capacity but enable governance integration by default.

Intervention and targeted modification. Mechanistic interpretability employs activation patching and causal tracing to experimentally manipulate intermediate activations, determining which components causally contribute to specific behaviors [7, 9, 8]. For example, by copying activations from a "clean" run into a corrupted context, one can isolate circuits or attention heads that restore correct outputs [9, 8]. This provides a falsifiable framework for testing internal hypotheses [4]. Beyond analysis, interpretability enables targeted intervention through circuit editing, head ablation, or representation reweighting to suppress undesired behaviors while preserving functionality [43, 44, 45]. Mechanistic insights can guide behavioral methods like RLHF, creating hybrid approaches that combine scalable alignment with causal guarantees [40, 37].

Detecting deceptive alignment and global reasoning. Interpretability provides defense against deceptive alignment by tracing goal-directed circuits and identifying reward hacking mechanisms [46, 47]. Toy models trained on mathematical tasks reveal symbolic operations, providing blueprints

Category	Limitation	Governance Impact
Representation	Polysemantic neurons, entangled features	Unclear regulatory compliance
Methodology	Post-hoc methods unstable/manipulable	Risk of "explanation theater"
Evaluation	Lack of standardized benchmarks	Weak causal validation for audits
Conceptual	Explanations diverge from human categories	Symbolic vs. subsymbolic mismatch
Practical	Compute- and expert-intensive	Limited accessibility and scalability
Risk Factors	Bias amplification, security leaks	Explanation laundering concerns

Table 2: Key limitations of interpretability methods and their governance implications.

for regulating reasoning in larger systems [48, 9, 8]. Causal mediation analysis extends interpretability from isolated units to entire pathways, offering global views of model reasoning [26, 44].

Governance integration. These capabilities support regulatory compliance: audit hooks enable regulators to trace decisions to internal circuits, circuit editing allows targeted mitigation of bias or toxicity without model redeployment, and provenance tracking satisfies documentation requirements under frameworks like the EU AI Act. By embedding interpretability as a design constraint rather than retrofitting it post-deployment, systems become auditable and governable by construction. This positions interpretability as a governance prerequisite for compliance with emerging regulatory frameworks.

Interpretability as a Private Oversight Mechanism. Interpretability-first architectures can underpin third-party audits, compliance certifications, and risk insurance models [14, 15]. Recent frameworks for private governance of frontier AI [14] and markets for AI governance [15] highlight the need for technical substrates that generate verifiable causal evidence. Accountability in algorithmic supply chains [49] requires mechanisms for tracing decisions across distributed systems. Mechanistic interpretability provides the technical substrate for private assurance—enabling certifiers to verify causal alignment claims through reproducible audits, much as insurers verify risk portfolios through actuarial evidence. This connects causal abstraction to "assurance evidence pipelines" where LoBOX and MIB benchmarks feed into private audit dashboards, enabling third-party oversight bodies to validate model behavior through intervention-based testing.

Technical implementation for private oversight. Private governance mechanisms require specific technical capabilities beyond conceptual frameworks. Audit hooks must generate immutable logs of model decisions with circuit-level attribution, enabling third-party auditors to trace specific outputs to internal computational pathways. For example, a model with modular attention heads could allow auditors to verify that bias detection circuits activate appropriately across demographic groups, providing evidence for anti-discrimination compliance. Certification bodies need reproducible testing protocols that validate alignment claims across model versions—mechanistic interpretability enables this through standardized intervention tests that confirm circuit behavior remains consistent. Insurance frameworks require risk assessment metrics based on mechanistic understanding of failure modes: insurers could analyze the robustness of safety-critical circuits, quantify polysemanticity in decision-making layers, and assess the presence of known deceptive alignment patterns. Procurement governance can specify auditable transparency requirements in vendor contracts, requiring models to demonstrate circuit-level traceability for high-stakes decisions. These capabilities transform interpretability from a research tool into operational infrastructure for private governance.

While the potential of interpretability-as-design is significant, it is important to acknowledge the substantial limitations and challenges that the field currently faces.

5 Limitations and Critiques of Interpretability

While interpretability is essential for AI safety and transparency, it faces significant limitations—technical, conceptual, and practical. These challenges impact both method reliability and broader epistemic claims, as in Table 2.

Representational and methodological challenges. Polysemanticity—individual neurons encoding multiple unrelated features—complicates semantic interpretation and becomes more pronounced at scale [8, 7]. This should be distinguished from superposition, which refers to basis non-orthogonality in representation spaces. Polysemanticity (representational entanglement) creates a scale mismatch

Table 3: Comparison of AI Alignment Approaches. MI = Mechanistic Interpretability, RLHF = Reinforcement Learning from Human Feedback, Const. AI = Constitutional AI.

Criterion	MI	RLHF	Red Teaming	Const. AI
Transparency	High (internal)	Low (outputs)	Low (failures only)	Medium (principles)
Human Supervision	Low (experts)	High (raters)	High (testers)	Medium (curation)
Scalability	Medium	Low	Low	High
Causal Understanding	Yes	No	No	No
Risk Coverage	Inner failures	Behavioral	Exploits	Norms

between human cognition and model reasoning, as interpretability tools are poorly equipped to track long-range dependencies or emergent behavior across thousands of layers [45]. Recent work by Meloux et al. [32] questions whether mechanistic interpretations are even identifiable, while Sutter et al. [33] argue that causal abstraction alone may be insufficient for mechanistic interpretability. Post-hoc methods rely on surrogate explanations rather than causal mechanisms, are easily manipulated, and often lack theoretical guarantees [17, 21, 23]. Without rigorous validation through counterfactual testing or causal probing, these methods risk offering illusions of understanding [45, 26].

Evaluation and benchmarking gaps. The field lacks standardized benchmarks for explanation quality, with quantitative metrics often focusing on proxies like sparsity rather than causal validity [50, 29, 41, 12]. Recent frameworks like xai_evals [51] provide systematic evaluation of post-hoc explanation methods, while the MIB benchmark [29] provides causal fidelity evaluation through circuit localization tasks. Feature consistency [28]—the stability of learned representations across contexts—emerges as a critical evaluation criterion. Without widespread adoption of such benchmarks and metrics, the field risks fragmentation and irreproducibility. Calls for "role-calibrated" and context-sensitive interpretability reflect the need to move beyond shallow heuristics toward explanations that serve specific epistemic and safety purposes [8, 29].

Ethical and epistemic constraints. Interpretability techniques assume internal features map to linguistic categories, but evidence suggests this mapping is frequently indirect [7, 45, 8]. Research is susceptible to confirmation bias and "explanation theater"—compelling but unfounded justifications [36, 12, 23, 26]. Security risks include proprietary information exposure, adversarial attacks, and explanation laundering [21, 11, 52]. The LoBOX framework [39] proposes bounded opacity—selective transparency calibrated to institutional roles—acknowledging that full transparency may be neither feasible nor desirable.

To better understand interpretability's role in the broader alignment landscape, we now compare it with other alignment approaches.

6 Interpretability and Private Oversight Pipelines

Interpretability plays a distinct role in AI alignment, targeting underlying mechanisms of model while behavioral methods focus on outputs. We compare these approaches to situate interpretability within the alignment toolkit.

Behavioral alignment methods. RLHF aligns behavior through reward modeling and policy optimization [2, 53], but primarily addresses surface-level alignment without verifying internal reasoning safety [54]. Models may exhibit inner misalignment—producing aligned outputs while pursuing misaligned objectives [55, 46, 38, 56]. Red teaming probes models adversarially to uncover vulnerabilities [57], but reveals *that* failures occur without explaining *why* [58]. Constitutional AI aligns models with normative principles through supervised fine-tuning [59, 60], but operates at the behavioral level without confirming internal structure. These approaches are often preferred for their scalability, generalizing across tasks and domains with limited model-specific insight [1, 54].

We compare major alignment approaches to situate interpretability within the broader landscape (Table 3).

Interpretability as complement. Interpretability offers tools for probing internal representations and identifying latent goals, deceptive heuristics, or emergent failure modes that behavioral methods may overlook [4, 45]. It can augment red teaming by analyzing internal mechanisms that make models

susceptible to attacks—reliance on ambiguous embeddings, exploit-prone circuits, or memorized failure patterns [3, 30, 45]. Unlike behavioral methods that rely on human-centered assessments, mechanistic interpretability prioritizes testable, manipulable, and causally valid explanations [61, 4, 62, 45].

Hybrid approaches and epistemic value. Behavioral methods can shape outputs at scale while interpretability verifies causal fidelity post hoc. Models tuned with RLHF or Constitutional AI can be examined with activation patching and mediation to detect reward hacking, deceptive alignment, or brittle circuits that pass surface tests [45, 7, 43, 44, 40, 9]. Unlike behavioral metrics emphasizing persuasiveness, interpretability evaluates causal correctness, providing the epistemic backbone for trustworthy alignment. This hybrid approach could complement traditional audits under the EU AI Act, offering causal guarantees beyond behavioral testing. As models become more powerful and autonomous, alignment strategies that rely solely on behavioral feedback will become increasingly insufficient.

Hybrid governance pipelines. Mechanistic audits can be layered atop RLHF training pipelines to provide both behavioral and causal validation. This approach addresses the concern that alignment faking and deceptive alignment pose significant risks for regulators, particularly in General Purpose AI (GPAI) systems. Recent work by Hilton [40] demonstrates how formal verification can be combined with heuristic explanations, creating hybrid systems that maintain both interpretability and performance. Such pipelines enable regulators to verify that behavioral improvements correspond to genuine internal alignment rather than surface-level optimization.

Beyond technical comparison, these approaches differ in their suitability for private governance mechanisms. Interpretability enables third-party verification through reproducible causal evidence, while behavioral methods like RLHF face sociotechnical limitations in generating auditable assurance evidence [63]. This positions interpretability as uniquely suited for private oversight contexts requiring independent verification, such as certification bodies validating alignment claims, insurers assessing risk through causal evidence, and procurement standards specifying auditable transparency requirements.

The implications of interpretability extend well beyond technical considerations, touching on fundamental questions about AI governance, human understanding, and the future of AI development.

7 Private Governance through Interpretability

Interpretability provides causal understanding, internal auditing, and structured intervention [4, 3, 8], intersecting with governance, philosophy, regulatory design, and public trust [40, 37, 45].

Governance and auditability. Mechanistic interpretability signals a transition from opaque optimization to cognitively structured, editable artifacts. In high-stakes applications—such as healthcare, legal reasoning, or autonomous systems—transparent reasoning processes enable causal attribution in failures, making responsibility assignment and due process possible [36, 40, 1]. For General Purpose AI (GPAI) systems, interpretability could reshape legal liability regimes and inform consent standards by grounding them in internal explanations rather than surface behavior.

Markets for assurance and certification. Private governance mechanisms create markets for AI assurance, where interpretability provides the technical substrate for verifiable evidence [15]. Certification bodies can validate alignment claims through reproducible causal evidence, while insurers can assess risk through mechanistic understanding of model behavior. Procurement governance [64] can specify auditable transparency requirements, enabling buyers to verify model capabilities through interpretability evidence. Data intermediaries [65] can leverage interpretability to provide responsible data stewardship, while evaluation infrastructure [66] can incorporate mechanistic insights into assessment frameworks.

Cognitive alignment and architecture trade-offs. Human explanations are symbolic and narratively coherent; neural explanations are distributed and subsymbolic [26, 36]. This challenges assumptions that model explanations align with human cognitive categories [3, 67]. Current architectures prioritize performance, resulting in entangled representations that resist decomposition [9, 68]. Models designed with interpretability constraints—modularity, sparsity, hierarchical structuring—may yield transparent representations without sacrificing capability, supporting interpretability-first training paradigms [4, 48].

Bridging research and practice. Significant gaps remain between research demonstrations and deployment-ready systems. Real-world governance requires tooling, training, and institutional capacity that don't exist at scale. Closing these gaps demands interdisciplinary collaboration treating interpretability as a governance prerequisite from system conception. In competitive AI development, actors may resist transparency due to intellectual property concerns, but opacity increases systemic risk [38, 59, 40, 12]. The development of third-party interpretability audits, publicly maintained benchmarks, and regulatory mandates could align economic incentives with safety goals—enabling transparency without requiring complete openness.

8 Conclusion

Interpretability is a foundation for building safe and reliable AI systems, and increasingly, a technical substrate for private AI governance. As frontier models proliferate, private mechanisms—audits, certification, insurance, procurement—emerge to complement public regulation [14, 15]. While behavioral alignment strategies like RLHF, Constitutional AI, and red teaming shape outputs to human preference [55, 60, 46], they face sociotechnical limitations in generating verifiable assurance evidence [63]. Interpretability addresses this gap by reaching into hidden computations to understand, verify, and intervene on model reasoning, providing the causal evidence required for private oversight [30, 9].

However, neural networks encode features in overlapping, distributed ways that don't cleanly map to human concepts. Explanations are challenging to scale, hard to validate, or misleading when they mirror expectations rather than reality [36, 17, 37, 11]. We also lack automated tools that can operate at the scale of today's largest models. Interpretability research must face these challenges directly, including the ethical and epistemological stakes of valid explanation [40, 36, 26]. Explanations must be grounded in causal evidence and open to scrutiny [61, 62, 41].

AI alignment requires a joint approach: using interpretability as a design principle to shape model construction [4, 30, 8], and applying behavioral methods to guide external performance [55, 60, 46]. The goal is ensuring models are internally structured in ways that are understandable, inspectable, and aligned with human intent [38, 40, 45]. Interpretability is not a nice-to-have but a requirement for building systems we can audit, trust, and control [3, 4, 26]. Without it, alignment becomes a matter of hope. With it, we have a path to reasoning about AI in terms we can understand and shape [38, 4, 1].

The governance triangle (Figure 2) illustrates how interpretability bridges mechanistic design, regulatory audits, and governance frameworks. This infrastructure enables private governance mechanisms to operate effectively: certification bodies can validate alignment claims, insurers can assess risk through causal evidence, and procurement standards can specify auditable transparency requirements.

Future directions. Realizing interpretability as private governance infrastructure requires developing auditable interpretability pipelines—end-to-end systems that generate, validate, and communicate assurance evidence to third-party oversight bodies. We propose pilot projects for interpretability audits on medium-scale open models as practical stepping stones. Realizing interpretability-as-design requires extending benchmarks like MIB [29] to governance-relevant tasks, developing prototype models with modularity and audit hooks, building open-source audit toolchains, piloting circuit-level audits with regulators, and establishing interdisciplinary teams for co-design. Metrics should emphasize measurable outcomes including audit coverage, feature consistency [28], interpretability stability, causal fidelity, intervention success rates, assurance evidence quality (reproducibility, causal validity), certification workflow efficiency, and interoperability with governance frameworks. These frameworks—MIB for empirical validation, LoBOX [39] for role-calibrated transparency—provide foundations for operationalizing interpretability-as-governance infrastructure.

9 Impact Statement

This position paper argues for interpretability as infrastructure for private AI governance. Potential benefits include enabling third-party audits through reproducible causal evidence, supporting certification frameworks with verifiable alignment claims, facilitating risk assessment for insurance mechanisms, and providing technical substrates for procurement governance standards. Risks include "explanation theater" undermining audit compliance and verification protocols, potential exposure of proprietary model internals creating competitive disadvantages, and concentration of

interpretability expertise in well-resourced organizations limiting accessibility of private governance mechanisms. We advocate causal validation, rigorous benchmarks, and integration with private governance frameworks to mitigate these risks.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. URL https://arxiv.org/abs/1606.06565.

[2] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

[3] Lingbai Kong, Wengen Li, Hanchen Yang, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. Causalformer: An interpretable transformer for temporal causal discovery, 2024. URL https://arxiv.org/abs/2406.16708.

[4] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024. URL https://arxiv.org/abs/2404.14082.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

[6] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

[7] 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.

[8] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

[9] Elhage et al. Mathematical frameworks for mechanistic interpretability. *Anthropic*, 2021.

[10] 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.

[11] 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.

[12] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL https://arxiv.org/abs/1702.08608.

[13] Andrew D Selbst, Julia Powles, Zachary C Lipton, Seeta Peña Gangadharan, and Kate Crawford. The intuitive appeal of explainable machines. *Fordham Law Review*, 87(3):1085–1139, 2018.

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

[15] Philip Moreira Tomei, Rupal Jain, and Matija Franklin. Ai governance through markets, 2025. URL https://arxiv.org/abs/2501.17755.

[16] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016. URL http://arxiv.org/abs/1606.03490.

[17] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019. URL https://arxiv.org/abs/1811.10154.

[18] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2018. URL https://arxiv.org/abs/1711.00399.

[19] Tristan Gomez and Harold Mouchère. Enhancing post-hoc explanation benchmark reliability for image classification, 2023. URL https://arxiv.org/abs/2311.17876.

[20] Joseph Carlsmith. Is power-seeking ai an existential risk?, 2024. URL https://arxiv.org/abs/2206.13353.

[21] Christoph Molnar. *Interpretable Machine Learning*. Leanpub, 2 edition, 2022. https://christophm.github.io/interpretable-ml-book/.

[22] Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kuehnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. Neural-symbolic learning and reasoning: A survey and interpretation, 2017. URL https://arxiv.org/abs/1711.03902.

[23] 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.

[24] 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.

[25] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2025. URL https://arxiv.org/abs/2301.04709.

[26] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[27] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic Research*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

[28] Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. Mechanistic interpretability should prioritize feature consistency in saes. OpenReview, 2025. URL https://openreview.net/forum?id=d9ACURK6bI. Published September 30, 2025.

[29] Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. Mib: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*, 2025.

[30] Nelson Elhage et al. A mathematical framework for transformer circuits. *Anthropic*, 2021.

[31] Lawrence Chan. Superposition is not "just" neuron polysemanticity. https://www.alignmentforum.org/posts/8EyCQKuWo6swZpagS/superposition-is-not-just-neuron-polysemanticity, 2024. AI Alignment Forum post.

[32] Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? *arXiv preprint arXiv:2502.20914*, 2025.

[33] Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? *arXiv preprint arXiv:2507.08802*, 2025.

[34] Ananya Joshi, Celia Cintas, and Skyler Speakman. Enabling precise topic alignment in large language models via sparse autoencoders. *arXiv preprint*, 2025.

[35] Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Reduan Achtibat, Patrick Kahardipraja, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. Attribution-guided pruning for compression, circuit discovery, and targeted correction in llms. *ArXiv*, abs/2506.13727, 2025. URL https://api.semanticscholar.org/CorpusID:279410755.

[36] Huiren Bai. The epistemology of machine explanations. *FILOSOFIJA. SOCIOLOGIJA*, 2022.

[37] Carla Capeto. Theatre and ai: A brief study of ethics in narratives and performance. *ScienceOpen*, 2024.

[38] Kcyras. Alignment gaps. *Lesswrong*, 2024.

[39] Fernando Herrera and Rodrigo Calderón. Opacity as a feature, not a flaw: The lobox governance ethic for role-sensitive explainability and institutional trust in ai, 2025.

[40] Jacob Hilton. Formal verification, heuristic explanations and surprise accounting. *AI Safety Journal*, 2024.

[41] Eduardo M. Pereira Diogo V. Carvalho and Jamie S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *MDPI*, 2019.

[42] Nikita Malik, Pratinav Seth, Neeraj Kumar Singh, Chintan Chitroda, and Vinay Kumar Sankarapu. Interpretability-aware pruning for efficient medical image analysis, 2025. URL https://arxiv.org/abs/2507.08330.

[43] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

[44] Xander Davies Maximilian Li and Max Nadeau. Circuit breaking: Removing model behaviors with targeted ablation. *arXiv*, 2024.

[45] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020. URL https://arxiv.org/abs/2004.12265.

[46] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2021. URL https://arxiv.org/abs/1906.01820.

[47] Priyanka Bharadwaj. Why eliminating deception won't align ai, 2025.

[48] Anh Tong et al. Neural ode transformers: Analyzing internal dynamics and adaptive fine-tuning. *ICLR Workshop*, 2024.

[49] Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1186–1197. Association for Computing Machinery, 2023.

[50] Pratinav Seth and Vinay Kumar Sankarapu. Bridging the gap in xai-why reliable metrics matter for explainability and compliance, 2025. URL https://arxiv.org/abs/2502.04695.

[51] Pratinav Seth, Yashwardhan Rathore, Neeraj Kumar Singh, Chintan Chitroda, and Vinay Kumar Sankarapu. xai_evals : A framework for evaluating post-hoc local explanation methods. *ArXiv*, abs/2502.03014, 2025. URL https://api.semanticscholar.org/CorpusID:276116864.

[52] Suraj Srinivas and François Fleuret. Gradient alignment in deep neural networks. *CoRR*, abs/2006.09128, 2020. URL https://arxiv.org/abs/2006.09128.

[53] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

[54] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018. URL https://arxiv.org/abs/1811.07871.

[55] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

[56] Casey Barkan, Sid Black, and Oliver Sourbut. Do llms know what they're capable of? why this matters for ai safety, 2025.

[57] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.

[58] Silvan Ferreira, Ivanovitch Silva, and Allan Martins. Organizing a society of language models: Structures and mechanisms for enhanced collective intelligence, 2024. URL https://arxiv.org/abs/2405.03825.

[59] Anthropic. The need for transparency in frontier ai. *Policy Report*, 2025.

[60] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

[61] Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts, 2023. URL https://arxiv.org/abs/2305.19951.

[62] Gwenn Englebienne Andrea Papenmeier, Dagmar Kern and Christin Seifert. It's complicated: The relationship between user trust, model accuracy and explanations in ai. *AMC Digital Libraries*, 2022.

[63] 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.

[64] Lavi M. Ben Dor and Cary Coglianese. Procurement as ai governance. *IEEE Transactions on Technology and Society*, 2(4):192–199, 2021. doi: 10.1109/TTS.2021.3111764.

[65] Jovan Powar, Heleen Janssen, Richard Cloete, and Jatinder Singh. From policy to practice in data governance and responsible data stewardship: system design for data intermediaries. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2491–2504. Association for Computing Machinery, 2025.

[66] Merlin Stein, Milan Gandhi, Theresa Kriecherbauer, Amin Oueslati, and Robert Trager. Who should run advanced ai evaluations – aisis?, 2025. URL https://arxiv.org/abs/2407.20847.

[67] Iwan Williams, Ninell Oldenburg, Ruchira Dhar, Joshua Hatherley, Constanza Fierro, Nina Rajcic, Sandrine R. Schiller, Filippos Stamatiou, and Anders Søgaard. Mechanistic interpretability needs philosophy. *arXiv preprint arXiv:2506.18852*, 2025.

[68] Satvik Golechha and James Dao. Challenges in mechanistically interpreting model representations. *arXiv preprint arXiv:2402.03855*, 2024.

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 interpretability should be treated as a design principle for alignment—and this is consistently argued across Sections 1–8 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 5 (*Limitations and Critiques of Interpretability*) systematically discusses technical, conceptual, and practical limitations, including scalability, polysemanticity, evaluation gaps, and risks of explanation theater.

   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 8).

   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 outlines both positive impacts (auditable AI, regulatory 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 referenced are properly cited with their original sources throughout the paper.

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.