Interpretability Can Be Actionable

Hadas Orgad [1]
Fazl Barez [*2 3] **Tal Haklay** [*4] **Isabelle Lee** [*5] **Marius Mosbach** [*6 7] **Anja Reusch** [*4] **Naomi Saphra** [*1 8]
Byron C Wallace [*9] **Sarah Wiegreffe** [*10] **Eric Wong** [*11]
Ian Tenney [12] **Mor Geva** [13]

Abstract

Interpretability aims to explain the behavior of deep neural networks. Despite rapid growth, there is mounting concern that much of this work has not translated into practical impact, raising questions about its relevance and utility. This position paper argues that the central missing ingredient is not new methods, but evaluation criteria: interpretability should be evaluated by *actionability*—the extent to which insights enable concrete decisions and interventions beyond interpretability research itself. We define actionable interpretability along two dimensions—concreteness and validation—and analyze the barriers currently preventing real-world impact. To address these barriers, we identify five domains where interpretability offers unique leverage and present a framework for actionable interpretability with evaluation criteria aligned with practical outcomes. Our goal is not to downplay exploratory research, but to establish actionability as a core objective of interpretability research.

Figure 1. Actionability checklist for interpretability research.

1. Introduction

Interpretability research seeks to explain modern machine learning systems. In recent years, it has grown into a large and active research area (Mosbach et al., 2024; Maslej et al., 2025), driven by the intuition that understanding models should help make them more reliable, efficient, safer, and aligned with human values (Bereska & Gavves, 2024).

Despite its growth, interpretability work is often seen as lacking practical impact such as informing changes to models, training practices, deployment decisions, or policy (Krishnan, 2020; Greenblatt et al., 2023; Potts, 2025), motivating calls to focus on clearly demonstrable outcomes beyond "understanding" itself (Haklay et al., 2025b; Upadhyay & Barez, 2025; Nanda et al., 2025; Barez, 2026). Our framing draws in part on discussions from the ICML 2025 work-

[*]Equal contribution [1]Kempner Institute at Harvard University [2]University of Oxford [3]Martian [4]Technion—IIT [5]University of Southern California [6]Mila – Quebec AI Institute [7]McGill University [8]Boston University [9]Northeastern University [10]University of Maryland [11]University of Pennsylvania [12]Google DeepMind [13]Tel Aviv University. Correspondence to: Hadas Orgad <hadasorgad@fas.harvard.edu>.

shop on actionable interpretability, which aimed to foster dialogue on leveraging interpretability insights to drive tangible advancements in AI.

In this paper, we argue that interpretability research should be evaluated not only by how well it explains models, but by what those explanations enable us to do. That is, interpretability should be held to a standard of *actionability*.

We contend that the field's impact will be strengthened if it explicitly tracks not only what we understand, but what that understanding enables us to do. We do not, however, argue that all interpretability work must immediately yield actionable outcomes, nor that purely exploratory contributions lack value. Indeed, methodological novelty and demonstrated applications are not at odds—grounding findings in real-world actions holds methods to a higher standard, providing evidence that insights reflect genuine model behavior rather than artifacts of a particular analysis. What is missing in interpretability research is not *methods*, but *evaluation criteria*: a shared framework for determining when interpretability research is successful from a practical, decision-oriented perspective. We therefore advance a framework for actionable interpretability: analyzing current limitations, identifying opportunities for impact, and suggesting practical tools to increase actionability. Figure 1 translates our central thesis into concrete steps for researchers.

Scope. We consider interpretability in modern machine learning, focusing on deep learning and foundation models. Although we draw many examples from LLMs, our arguments apply broadly to any deep neural networks domain which may require explanation. This is a position paper: rather than an exhaustive survey, we propose actionability as a unifying lens for evaluating interpretability work.

The paper is organized as follows: Section 2 defines actionable interpretability. Section 3 diagnoses the barriers that currently prevent interpretability from achieving real-world impact. The rest of the paper paves the way towards more actionable interpretability. Section 4 identifies opportunities for actionability. Section 5 presents a framework for categorizing actions and Section 6 discusses evaluation criteria aligned with actionability. Section 7 addresses counter-arguments. Section 8 reviews related work. Section 9 concludes with an actionable checklist for researchers, summarized in Figure 1.

2. Defining Actionable Interpretability

We consider a work[1] to be *interpretability-oriented* if it aims to explain or analyze an AI model—for example, works that analyze model representations, explain specific behaviors or capabilities, or discover internal mechanisms. Having this distinction, we provide the following definition:

Actionable Interpretability An interpretability-oriented work is actionable if it produces *insights* about an AI model that inform or guide *actions* toward non-interpretability objectives.

Insights are outputs of interpretability work: findings about how models represent or process inputs, explanations of internal mechanisms, or methods that clarify behavior.

Actions (toward non-interpretability objectives) are decisions made by humans in response to interpretability insights that would not have been taken otherwise. These fall outside the scope of interpretability itself and ideally lead to concrete improvements such as enhanced performance, better-calibrated trust, or improved safety.

2.1. Dimensions of Actionability

In practice, actionability is more fine-grained and not binary. Interpretability-oriented work can support different levels of actionability, which we characterize along two key dimensions: *concreteness* and *validation*.

Concreteness captures how precisely an action is articulated. At the low end are vague suggestions ("could inform safety research") or no suggestions at all; at the high end are exact specifications with implementation details.

Validation captures empirical support for an action's utility. At the low end, actions are untested hypotheses; at the high end, they are systematically evaluated with quantitative or qualitative evidence of meaningful outcomes beyond interpretability research itself.

Together, these dimensions span a space for situating interpretability work (illustrated in Figure 3 in the Appendix):

Low concreteness, low validation: Work in this region recommends no specific actions to validate. The insights from this work may, however, inform future work by providing a starting point that others can build upon and test. For example, Geva et al. (2021)'s key-value memory view of MLPs directionally motivated subsequent work on knowledge localization and model editing. Wang et al. (2023), Conmy et al. (2023) and others laid groundwork for circuit-based analysis. While not the emphasis of this paper, such exploratory work is imperative to drive the field forward.

High concreteness, low validation: Concrete actions proposed but not empirically validated—e.g., approaches for verifying scientific models to build trust in their predictions (King et al., 2025; Li et al.; Ferreira et al., 2025) or optimizing model deployment and training (Zhao et al., 2025; Chen et al., 2025).

[1] By "work", we refer broadly to research or engineering contributions, including methods, models, analyses, benchmarks, and empirical studies.

High concreteness, high validation. Precise specifications with demonstrated utility, informed by interpretability insights drawn either from the work itself or prior work. Examples include model editing methods leveraging the MLP key-value store view (Meng et al., 2022; Wang et al., 2023; Arad et al., 2024; Fang et al., 2025), are based on sparse-auto-encoders (Gur-Arieh et al., 2025; Ashuach et al., 2025a) or insights into the role of cross-attention layers (Orgad et al., 2023; Gandikota et al., 2024). Representation finetuning (Wu et al., 2024), an alternative to LoRA-based methods, was inspired by interpretability findings. Schut et al. (2025) use concept vectors to uncover novel chess concepts transferable to human players. Anthropic (2025) analyzed internal activations during a safety audit of Claude.

3. Why Interpretability Isn't (Yet) Actionable

Despite growing interest, several barriers limit interpretability's real-world impact: misaligned incentives, methodological limitations, and deployment challenges. These reinforce a cycle where actionability is not prioritized, methods lack validation, and deployment yields little feedback. The rest of the paper discusses how to advance actionable interpetability despite these limitations

3.1. Misaligned Incentives

The interpretability community does not sufficiently reward work for demonstrating practical value. Without a strong incentive to prove that interpretability methods deliver real-world value, researchers are less likely to conduct or show interest in actionable interpretability work.

Publication standards do not require actionability. Papers can be accepted based purely on methodological novelty, with no requirement to demonstrate applications. Meanwhile, **application-focused work is under-rewarded,** Practical demonstrations may be dismissed as "merely engineering" despite their greater potential impact. We argue that methodological novelty and application demonstration are not at odds—demonstrating applications holds interpretability methods to a higher standard, providing evidence that findings are grounded in reality. This asymmetry—low requirements for actionability combined with low rewards for demonstrating it—substantially reduces researchers' incentive to pursue practical applications.

These issues are not unique to the interpretability field, and also exist in mainstream machine learning (ML) research. However, unlike applied ML, where benchmark performance provides immediate feedback, **interpretability lacks clear signals of success**. Mainstream ML research has a forcing function interpretability lacks: new methods must demonstrate gains on established benchmarks. The field has matured by moving from toy problems to real-world tasks—MNIST to ImageNet, Penn Treebank to diverse downstream tasks. However, the interpretability field has yet to fully mature, lacking agreed-upon standards.

3.2. Methodological Limitations

These incentive gaps often manifest as concrete technical problems that prevent interpretability insights from translating into action. In this section, we outline such technical problems and associated methods.

Lack of actionable insights. Interpretability work often fails to articulate how findings can inspire concrete actions. This limitation was reflected in the ICML 2025 workshop on Actionable Interpretability, where in 21.8% of the submitted papers, at least one reviewer explicitly flagged the work as insufficiently actionable. Mosbach et al. (2024) showed that although interpretability papers are cited, their impact is predominantly conceptual—most citations do not credit changes to training, architecture, or evaluation. While foundational work may eventually drive actionability (Bau, 2025), the field should explicitly reflect on how insights matter beyond its boundaries.

Oversimplified setups. Much research uses simplified tasks and small models. For instance, many mechanistic studies on LLMs focus on single next-token predictions (Mueller et al., 2025), whereas real usage involves multi-token generation. These settings are valuable as controlled testbeds, but their insights may not transfer to realistic settings. Recent work by Haklay et al. (2025a) has begun addressing these limitations with circuit discovery that handles variable-length inputs.

Insufficient comparative analysis. Many works lack rigorous comparisons against alternative approaches and fail to evaluate robustness across architectures, datasets, and tasks. As Casper (2023) argues, weak evaluation hinders progress toward practical tools. Recent benchmarks have begun to address this limitation, highlighting the importance of empirical comparisons. AxBench (Wu et al., 2025) showed that prompting and finetuning often outperform interpretability methods for LLM steering. MIB (Mueller et al., 2025) evaluates both circuit localization and causal variable localization—two widely studied directions that previously lacked a means to compare methods.

3.3. Deployment Challenges

Even when interpretability methods offer practical value, several barriers hinder their adoption.

Technical complexity. To employ interpretability techniques, a user must deeply understand model internals and be familiar with specialized libraries (Nanda & Bloom, 2022; Fiotto-Kaufman et al., 2025). Those outside the community often lack the expertise required (Ashtari et al., 2023)

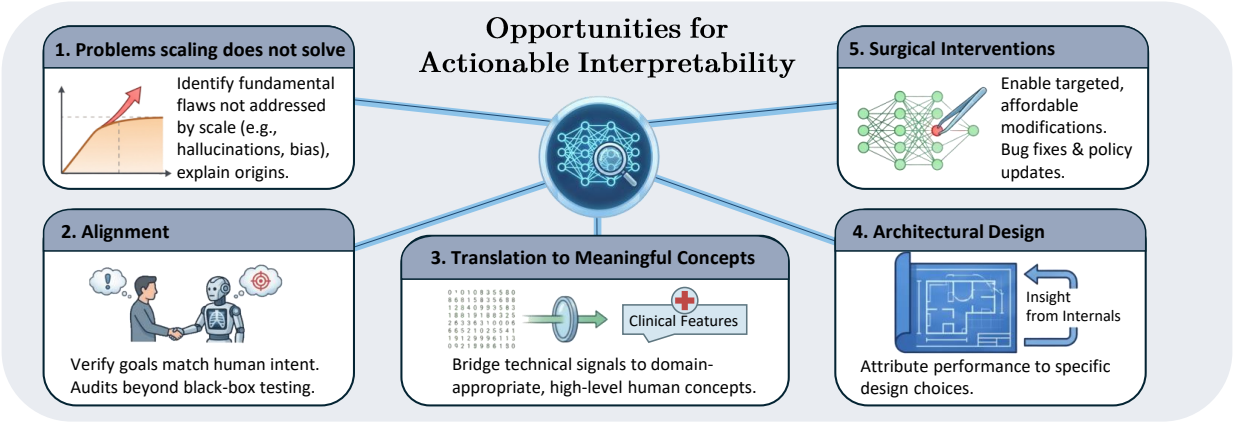

Figure 2. Five domains where interpretability offers unique leverage to drive concrete improvements.

and so rarely adopt these methods, especially when simpler alternatives exist.

The open-weights assumption. Most methods require direct access to weights and activations, restricting applicability to open-weight models. This creates a tension: interpretability is often motivated by safety concerns around powerful frontier models, yet these same models are often proprietary and therefore resistant to such analysis.

4. Making Interpretability Actionable

In Section 3, we identified limitations that prevent interpretability research from delivering sustained practical impact. Here, we turn to solutions: we identify opportunities where interpretability is uniquely positioned to drive concrete improvements. We identify five domains in which interpretability offers unique leverage—where there is a fundamental advantage from answering *why* questions about the model. In Sections 5 and 6 we discuss the implementation: a framework for actionable interpretability, and its evaluation.

Problems scaling does not solve. The scaling hypothesis—the claim that many capabilities improve predictably with increased model size—has proven remarkably successful. Yet certain failure modes persist or even worsen with scale, including hallucinations, catastrophic forgetting, biases and adversarial brittleness. The persistence of these failures across model scales suggests they are fundamental to our current modeling paradigm rather than due to limited capacity. Interpretability offers a path forward precisely because it can identify why models fail. Standard evaluations detect failures but cannot explain their origins or suggest principled interventions. By contrast, interpretability enables sharper hypotheses about underlying mechanisms and reasoning about potential fixes. Even partial insights can rule out hypotheses and guide the design of solutions.

Alignment. As AI systems become more capable, ensuring they behave as intended becomes more critical and more difficult. Alignment today still relies on fine-tuning and data curation rather than understanding-driven interventions, but as AI progresses, verifying that AI goals match human goals will shift from aspiration to necessity. Can we credibly claim a model has no deceptive capabilities without understanding its decision-making? Can we audit for backdoors through black-box testing alone? Since alignment concerns what a model optimizes for, it cannot be fully established without interpretability.

Surgical interventions. Retraining a flawed model is expensive and risks introducing other unexpected outcomes. Interpretability enables targeted modifications; identifying components responsible for unwanted behaviors allows surgical fixes while preserving other functionality. Though not yet fully practical, this is among interpretability research's most actionable outcomes—it's efficient and affordable. These techniques can enable post-hoc maintenance: bug fixes, policy updates, and rapid responses to new failure modes.

Architectural design. Current improvements emerge largely through trial and error—an inefficient, opaque process where success may not scale or transfer to new domains. Interpretability can transform this paradigm by linking specific design choices (data curation, architecture, optimization) to their effects on model behavior. This approach could accelerate progress by narrowing the space of plausible architecture modifications, reducing both labor and compute required.

Translation of explanation to meaningful concepts. The most natural role of interpretability is explaining model behavior, yet translating internal signals into meaningful concepts remains a critical bottleneck. In high-stakes domains like healthcare, a radiologist needs to know if an AI-assisted diagnosis depends on clinically relevant features, not which pixels activate; A developer debugging failures needs spe-

cific, legible insights than what current circuit discovery methods provide ("layers 7 and 9 interact together"). Automated methods that translate technical explanations into domain-appropriate, *actionable* concepts could unlock interpretability's core promise. This also includes methods that scale interpretability methods beyond a single input or template into a more natural, diverse setting.

5. A Framework for Actionable Interpretability

We now present a framework for the actions interpretability enables and the actors who carry them out. This framework is intended to help researchers identify and articulate the actionability potential of their own work. The examples we present throughout this section demonstrate successful cases of actionable interpretability, yet they represent a relatively small fraction of the broader literature.

Who takes the "action" in "actionable"? Different stakeholders have different capabilities and motivation, as illustrated in Table 1. **AI developers** may use mechanistic insights to inform model design. **Deployment engineers** may focus on controlling behavior in specific applications. **Domain experts** like clinicians need feature-level rationales to justify diagnoses. **A policymaker** relies on system-level summaries of fairness or compliance. These actors rarely operate in isolation—interpretability outputs should serve as communication interfaces across roles. A clinician's feedback about unreliable explanations may reveal failure modes to engineers. Similarly, a policy maker's compliance requirements may drive developers toward mitigation.

An interpretability work will become more actionable *if it is explicit about its intended audience and the decisions it aims to support.* At the same time, Section 3.3 emphasizes technical complexity as a major barrier to deployment. Taken together, these observations suggest that effective actionable interpretability work must do more than produce insights or analyses: it must specify who can act on its findings, what actions are enabled and how (by providing code or explicit instructions), and where those actions plausibly apply.

We next describe the different types of actions that interpretability insights can drive. We classify them according to what each action affects. Rather than providing an exhaustive taxonomy, we provide representative examples for each. Additional examples are listed in Appendix B.

5.1. Actions that Modify Model Output

Interpretability can inform decisions that directly modify model behavior—changes in training, inputs, weights, or internal computations. These decisions are primarily made by developers and researchers with access to model internals.

Data curation. Influence functions can help identify training examples that help or harm model performance. Koh & Liang (2017) used them to detect mislabeled examples and improve accuracy. Han et al. (2020) used them to expose artifacts in training data. More recently, Agia et al. (2025) applied influence functions to robot learning, identifying detrimental demonstrations and achieving state-of-the-art results with only 33% of the original training data.

Model input. Interpretability can inform decisions about what inputs to provide to models. Zhou et al. (2024) built on the insight that transformers implement an internal optimizer for in-context learning (Akyürek et al., 2023), using influence functions to identify which in-context demonstrations help versus harm performance.

Training decisions. Casper et al. (2024a) built on insights about how models internally represent concepts and used latent adversarial training—perturbing internal representations—to defend against unforeseen vulnerabilities, thereby removing backdoors and improving robustness.

Direct control. Interpretability can identify components responsible for specific behaviors, enabling targeted interventions. Model editing modifies weights to insert, remove, or correct behaviors without full retraining (Meng et al., 2022; 2023; Orgad et al., 2023). Runtime interventions steer activations along interpretable directions at inference time (Li et al., 2023; Turner et al., 2023). Concept bottleneck models introduce human-defined concepts as intermediate representations, enabling expert-guided control (Koh et al., 2020b; Yuksekgonul et al., 2023; Oikarinen et al., 2023).

Safety. Interpretability can support efforts to remove or suppress unsafe behaviors encoded in model weights. Techniques such as concept erasure (Ravfogel et al., 2020; Elazar et al., 2021) and machine unlearning (Gandikota et al., 2024; Ashuach et al., 2025b; Bourtoule et al., 2021; Cao & Yang, 2015) provide principled approaches for mitigating privacy risks and removing unwanted behaviors by identifying and neutralizing specific learned associations.

5.2. Actions about Deployment and Use

Interpretability can inform decisions that end users—including domain experts (e.g., clinicians) and other practitioners—make when interacting with model outputs. Unlike decisions that modify the model's final output, these actions change *what humans do* with model predictions: when to trust them, when to override them, and how to integrate them into their workflows.

End user decisions. Prenosil et al. (2025) developed a neuro-symbolic system combining GPT-4 with rule-based expert systems for clinical data extraction, providing the transparency and auditability that enabled radiologists to confidently use AI while maintaining oversight. Activa-

Table 1. Different audiences of interpretability and their actionable outputs.

Audience	Example Action	Desired Output Type
AI developers and researchers	Curate training data; remove or enhance specific behaviors	Data-point level analysis; behavior modification methods
AI deployment engineers	Debug application-specific failures	Explanations for model errors
Domain experts and practitioners	Validate reasoning; refine workflows	Explanations tied to domain features
End users	Trust or override output, adjust behavior	High-level rationale in human terms, UX/UI for steering model behavior
Policymakers and auditors	Enforce compliance or transparency	System-level summaries

tion patching may reveal when models are (overly) relying on patient demographic information when making clinical predictions (Ahsan et al., 2025).

Work on uncertainty estimation from internal representations (Kadavath et al., 2022, inter alia) enables users to detect potential errors and make informed decisions about when to trust model outputs.

Deployment decisions. Interpretability enables predictions about where models will fail. Huang et al. (2025) use internal mechanisms to identify out-of-distribution failures during inference, while Li et al. (2025) use them to predict errors on unseen distributions. Such insights support routing decisions—whether to return a model's answer or escalate to alternative methods. Work that detects model errors based on internal representations (Kadavath et al., 2022, inter alia) can also be used in this context. Chen et al. (2024a) demonstrate the potential value, achieving up to 98% cost reduction while matching GPT-4 performance through uncertainty-based routing.

5.3. Shaping Future Practice

Beyond immediate interventions, interpretability may offer insights that inform how the field builds and governs future systems. This has longer-term, broader impact.

Policy and regulation. Interpretability requirements are increasingly embedded in regulations. The EU AI Act mandates explainability for high-risk AI systems (ArtificialIntelligenceAct.eu, 2025). GDPR's Article 22 (European Union, 2016) restricts automated decision-making with significant effects and requires safeguards, including the right for human intervention. A central challenge to regulation is verification: whether interpretability enables credible claims about the absence of dangerous mechanisms, even with limited access to proprietary model details. This is currently largely unsolved, even for public or open-weight models.

Learning from superhuman models. When models exceed human expertise, interpretability becomes a mechanism not just for trust or safety, but for transferring knowledge from machines back to humans. Schut et al. (2025) show that interpreting AlphaZero's (Silver et al., 2018) strategies surfaces novel chess concepts that can teach human grandmasters—demonstrating interpretability's potential for knowledge transfer from AI to humans.

Development of future models. Interpretability can shift model design from trial-and-error toward principled engineering guided by an understanding of the computations performed. For example, induction heads in Transformers (Elhage et al., 2021; Olsson et al., 2022) provided a mechanism for in-context learning that traditional state-space models lacked, directly influencing the design of the Mamba architecture's selective state updates (Gu & Dao, 2024).

6. Evaluating Actionability

How do we know if interpretability work is actionable? Current practice often evaluates methods against other interpretability techniques or relies on intuitive notions of "understanding". This is insufficient for actionable interpretability. Here, we require metrics that measure whether insights actually enable better decisions and outcomes. We propose evaluation criteria for insights that can enable each of the three action categories from Section 5. All of these criteria share a common principle: *interacting with the world beyond the field of interpretability* rather than solely comparing methods within the field.

6.1. Evaluating Actions that Modify Outputs

Comparative utility against standard baselines. A major limitation of current evaluation is "grading on a curve"—comparing interpretability methods only against each other. Instead, performance should be measured against standard, pragmatic ML baselines, such as prompting or fine-tuning, and using standard metrics such as accuracy, or benchmark-specific measures. For example, does steering with SAEs improve refusal behavior more than targeted prompting or small LoRA (Hu et al., 2022) adapters? This defines action-

ability as marginal leverage gained over simpler methods that do not require deep mechanistic understanding.

Mechanistic faithfulness. This measures whether an explanation correctly identifies model components causally involved in a specific computation. Evaluation uses intervention-based verification on tasks with well-defined semantics: an explanation is faithful if intervening on identified components produces predicted changes—altering target computations while leaving unrelated behaviors intact. For example, when reverse engineering an LLM's sorting algorithm, one can identify the comparison component and intervene to reliably swap two specific items in the output.

Generalization. To address whether an insight holds beyond a specific setting, our metrics must evaluate whether it generalizes, e.g., across seeds, input perturbations, architectures, and scales, without requiring rediscovery. Concrete evaluations may include transferring identified circuits, features, or mechanisms between models of different sizes, or from toy settings to frontier models. Successful transfer indicates the method captures robust, reusable structure.

Specificity. Next, we consider whether an interpretability claim identifies a component that is specifically linked to a distinctive target, rather than a broad correlation. This is evaluated in two ways. First, does the proposed component explain the behavior better than alternatives? This establishes that the finding is genuinely informative rather than arbitrary. Second, when intervening on the component to modify target behavior, do unrelated behaviors remain unchanged? This should be evaluated on standard benchmarks that quantify model capabilities. For example, if a neuron *specifically* controls review sentiment, an intervention may affect the tone while preserving the factual content and performance on unrelated tasks. Interventions that reveal broad effects suggest the component plays a generalized, entangled role in model behavior.

6.2. Evaluating Actions about Deployment

Task-enhancement. The most direct user-facing metric is whether explanations improve performance on the task the model supports—not the model itself, but human decision-making, speed, or reliance on outputs. This typically requires human-subject evaluations, which are critical since prior work suggests explanations do not reliably improve performance (Spillner et al., 2025).

Understandability. Incomprehensible explanations are unlikely to influence user decisions, even if technically correct. This is especially pronounced in high-stakes settings where practitioners face severe penalties for errors. For example, in medicine, clinical training requirements and legal liability create high barriers to AI adoption, even for superhuman systems. Importantly, understandability is orthogonal to faithfulness—an explanation may accurately reflect model behavior while failing to be usable. We expand on these evaluations in Appendix C.1.

Reliability. Even when explanations improve task performance and are understandable, they may fail to be actionable if perceived as unreliable. Explanations that vary substantially across random seeds or minor perturbations introduce uncertainty that undermines trust. While related to generalization (Cf. Section 6.1), reliability focuses on within-task stability—whether a user can expect consistent explanations across repeated or slightly varied contexts. This framing is especially important in high-stakes domains where explanations guide interventions and brittle explanations are often viewed as unsafe or uninformative (Ghassemi et al., 2021; Arun et al., 2021). We provide examples for measuring reliability in Appendix C.2.

6.3. Evaluating Actions Shaping Future Practices

In policy contexts, interpretability acts as an institutional lever rather than a scientific diagnostic. Actionability should be measured by whether methods enable feasible AI governance by regulators and safety teams (Upadhyay & Barez, 2025), and not by depth of insight for a handful of researchers staring at neuron visualizations. From this perspective, interpretability is policy-actionable to the extent that it expands feasible governance interventions: supporting safety audits, interpretable proxy models, or verifying the absence of dangerous mechanisms. Practically, interpretability should reduce monitoring and mitigation costs relative to blunt instruments like pausing deployment, supporting concrete policy tools (e.g., risk audits, model cards, licensing regimes) while remaining legible to non-experts.

7. Alternative Viewpoints

Is actionability the right goal? Some defend interpretability as basic science regardless of actionability. Bau (2025) argues for curiosity-driven research since we do not yet know which techniques may permit actionable insights. We do not disagree—we argue that impact will increase by tracking actionability as a yardstick, not that all work must be (immediately) actionable.

How should actionability be defined and measured? Many believe interpretability's value lies in AI safety (Nanda, 2022; Olah, 2023; Anwar et al., 2024; Amodei, 2025; Marks et al., 2025; Shah et al., 2025). Some argue safety is the *singular* actionable goal (Greenblatt et al., 2023; Nanda, 2022; Nanda et al., 2025; Hendrycks & Hiscott, 2025; Marks, 2025), and performance improvements represent "dual use" problem (Segerie, 2023; So8res, 2023; Shovelain & Mckernon, 2023). We argue for a broader framework centered on human users, encompassing both

safety and performance improvements.

Is actionability achievable? For practitioners whose priority is building better models, there must be decisive evidence that interpretability methods outperform alternatives with minimal additional effort. We currently lack this evidence for many research lines, contributing to distrust within the broader ML community. However, we can reduce skepticism by ensuring baselines include non-interpretability methods, as we discuss in Section 6. Recent community efforts aim to unify the discussion around actionability (Haklay et al., 2025b). Many remain optimistic, and the community is actively pivoting (Gao, 2025; Ho, 2025; Steinhardt & Schwettmann, 2024; Marks, 2025). We have additionally laid out arguments in this paper for why interpretability *is already actionable* in many scenarios (Section 5).

It would be premature to discount actionable interpretability when the field is still at an early-stage compared to to other scientific disciplines, and our objects of study have only emerged in their current form in the past five years.

8. Previous Work

Previous conceptual and position work. Lipton (2018) pointed out that interpretability is an overloaded term, and distinguishes between *transparency* (understanding how a model works) and *post-hoc interpretability* (explaining its decisions after the fact). Miller (2019) argues that interpretability requires attention to user and social context because much work neglects decades of findings from from philosophy, psychology, and cognitive science research which highlight how explanations should be grounded. Similarly, Jacovi & Goldberg (2021) emphasize the role of *social attribution* in explanations, namely the implicit attribution of intent to models. Rudin (2019) advances this perspective, suggesting that researchers should abandon post-hoc explanations of models entirely and instead focus on inherently interpretable models.

Others (Doshi-Velez & Kim, 2017) argue that without clear criteria, interpretability research may prioritize intuitively appealing methods over practically valuable ones. While their emphasis aligns with actionable interpretability, we contend evaluation should focus on the specific interventions and decisions the insights enable. Calderon & Reichart (2025) note that NLP interpretability often fail to generalize beyond their initial domains and stressed the importance of defining stakeholders. While complementing to our view, our focus is on translating insights into actionable outcomes.

Most recently, Nanda et al. (2025) advocate a "pragmatic" approach to interpretability that focuses on solving specific problems rather than solely reverse-engineering models, using meaningful "proxy tasks" to drive rapid iteration. On the other hand, Bau (2025) argues for the importance of curiosity-driven research, noting that we cannot yet predict which interpretability techniques may yield actionable insights in the future.

Actionable Explainability. The field of actionable explainability originates primarily from human-AI interaction (HCI) and algorithmic recourse research, focusing on enabling individuals to act on model outputs. An explanation is considered "actionable" if it helps a person understand the changes needed to receive a different outcome in the future (Joshi et al., 2019; Ustun et al., 2019; Karimi et al., 2021; Singh et al., 2024). For instance, increasing income to obtain a loan approval. Other approaches (Singh et al., 2023; Poyiadzi et al., 2020) define actionability as the ability to translate explanations into feasible behavioral changes, while the idea was also extended to human-in-the-loop settings (Saranti et al., 2022), allowing domain experts to directly adjust model parameters. While actionable explainability centers on enabling human action, actionable interpretability reframes interpretability as a way to also drive concrete improvements in model performance and reliability, not only human understanding.

9. Conclusion

In this position paper, we argue that interpretability can have greater real-world impact if actionability is incorporated as a core evaluation criterion. This is not to say that conceptual or theoretical work without immediate practical utility has no place in the field; such research remains valuable and necessary. Rather, making actionability a common evaluation criterion and explicitly tracking what insights make possible can accelerate progress for exploratory work.

We conclude by offering an actionable checklist for interpretability researchers.

1. **Define a clear goal.** Identify a specific problem that your interpretability question aims to eventually solve.
2. **Identify your audience.** Although academic papers are primarily read by researchers, their insights may be acted upon by different stakeholders (e.g., developers, practitioners, policymakers), each of whom may require different framing, language, or levels of abstraction.
3. **Propose concrete actions.** Articulate what decisions or interventions your insights enable.
4. **Validate empirically.** Where possible, implement the proposed action yourself and demonstrate its effects.
5. **Evaluate in realistic settings.** Apply your methods to realistic scenarios, including large-scale models and non-synthetic datasets.

6. **Use actionable metrics**, as descibed in Section 6. Especially, ask whether your contribution:
   - Surpasses standard baselines (e.g., prompting, finetuning) on target metrics.
   - Generalizes across models and other variations of the setting.
   - Produces targeted effects without degrading unrelated capabilities.
   - Yields explanations that are useful for the target audience—and if not, whether they can be translated into a more accessible form.

The burden now falls on the research community: to reward actionable contributions alongside explanatory depth, to establish evaluation criteria that track the utility of interpretability insights, and to build infrastructure that connects understanding to impact.

Acknowledgments

This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. A.R. was funded through the Azrieli international postdoctoral fellowship and the Ali Kaufman postdoctoral fellowship. B.W. is supported by a grant from Coefficient Giving and the National Science Foundation (NSF), RI 2211954. I.L. is supported by a Technical AI Safety research grant from Coefficient Giving via Berkeley Existential Risk Initiative. M.M. is supported by the Mila P2v5 grant and the Mila-Samsung grant. E.W. is supported by a grant from the National Science Foundation (NSF), CCF 2442421.

References

Achara, A., Anton, E. P., Hammers, A., and King, A. P. Invisible attributes, visible biases: Exploring demographic shortcuts in mri-based alzheimer's disease classification. *arXiv preprint arXiv:2509.09558*, 2025.

Agarwal, C., Tanneru, S. H., and Lakkaraju, H. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. Neural additive models: Interpretable machine learning with neural nets. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Agia, C., Sinha, R., Yang, J., Antonova, R., Pavone, M., Nishimura, H., Itkina, M., and Bohg, J. Cupid: Curating data your robot loves with influence functions. In Lim, J., Song, S., and Park, H.-W. (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2907–2932. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/agia25a.html.

Ahsan, H., Sharma, A. S., Amir, S., Bau, D., and Wallace, B. C. Elucidating Mechanisms of Demographic Bias in LLMs for Healthcare. In *Proceedings of the Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Amodei, D. The urgency of interpretability, April 2025. URL https://www.darioamodei.com/post/the-urgency-of-interpretability. Blogpost.

Anthropic. System card: Claude sonnet 4.5, 2025. https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., et al. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. doi: 10.48550/arXiv.2404.09932. URL http://arxiv.org/abs/2404.09932. arXiv:2404.09932 [cs].

Arad, D., Orgad, H., and Belinkov, Y. ReFACT: Updating text-to-image models by editing the text encoder. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2537–2558, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.140. URL https://aclanthology.org/2024.naacl-long.140/.

ArtificialIntelligenceAct.eu. Article 86: Right to explanation of individual decision-making. Online, 2025. URL https://artificialintelligenceact.eu/article/86/. Accessed: 2025-12-28.

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

Ashtari, N., Mullins, R., Qian, C., Wexler, J., Tenney, I., and Pushkarna, M. From discovery to adoption: Understanding the ml practitioners' interpretability journey. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pp. 2304–2325, 2023.

Ashuach, T., Arad, D., Mueller, A., Tutek, M., and Belinkov, Y. Crisp: Persistent concept unlearning via sparse autoencoders. *arXiv preprint arXiv:2508.13650*, 2025a.

Ashuach, T., Tutek, M., and Belinkov, Y. Revs: Unlearning sensitive information in language models via rank editing in the vocabulary space, 2025b. URL https://arxiv.org/abs/2406.09325.

Azaria, A. and Mitchell, T. The internal state of an LLM knows when it's lying. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Barez, F. Automated interpretability-driven model auditing and control: A research agenda. *AI Governance Initiative, University of Oxford*, January 2026. URL https://aigi.ox.ac.uk/publications/automated-interpretability-driven-model-auditing-and-control-a-research-agenda/. Working paper. Research agenda dated January 8, 2026.

Bassan, S. and Katz, G. Towards formal xai: formally approximate minimal explanations of neural networks. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pp. 187–207. Springer, 2023.

Bau, D. In defense of curiosity. https://davidbau.com/archives/2025/12/09/in_defense_of_curiosity.html, 2025. Accessed: 2025-12-09.

Bereska, L. and Gavves, S. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

Blanc, G., Lange, J., and Tan, L.-Y. Provably efficient, succinct, and precise explanations. *Advances in Neural Information Processing Systems*, 34:6129–6141, 2021.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *Proceedings of the 42nd IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

Brendel, W. and Bethge, M. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019.

Calderon, N. and Reichart, R. On behalf of the stakeholders: Trends in NLP model interpretability in the era of LLMs. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 656–693, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.29. URL https://aclanthology.org/2025.naacl-long.29/.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35. URL https://doi.org/10.1109/SP.2015.35.

Casper, S. Broad critiques of interpretability research, February 2023. URL https://www.lesswrong.com/posts/gwG9uqw255gafjYN4/eis-iii-broad-critiques-of-interpretability-research. The Engineer's Interpretability Sequence (blogpost 3 of 12).

Casper, S., Schulze, L., Patel, O., and Hadfield-Menell, D. Defending against unforeseen failure modes with latent adversarial training. *Transactions on Machine Learning Research*, 2024a. URL https://openreview.net/forum?id=mVPPhQ8cAd.

Casper, S., Yun, J., Baek, J., Jung, Y., Kim, M., Kwon, K., Park, S., Moore, H., Shriver, D., Connor, M., et al. The satml'24 cnn interpretability competition: New innovations for concept-level interpretability. *arXiv preprint arXiv:2404.02949*, 2024b.

Chakraborti, T., Banerji, C. R., Marandon, A., Hellon, V., Mitra, R., Lehmann, B., Bräuninger, L., McGough, S., Turkay, C., Frangi, A. F., et al. Personalized uncertainty quantification in artificial intelligence. *Nature Machine Intelligence*, 7(4):522–530, 2025.

Chen, L., Zaharia, M., and Zou, J. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R. Featured Certification.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Chen, Y., Wu, A., DePodesta, T., Yeh, C., Li, K., Marin, N. C., Patel, O., Riecke, J., Raval, S., Seow, O., Wattenberg, M., and Viégas, F. Designing a dashboard for transparency and control of conversational ai, 2024b. URL https://arxiv.org/abs/2406.07882.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Crabbé, J. and van der Schaar, M. Evaluating the robustness of interpretability methods through explanation invariance and equivariance. *Advances in Neural Information Processing Systems*, 36:71393–71429, 2023.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. Amnesic probing: Behavioral explanation with amnesic counterfactuals. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 9, pp. 147–163, 2021. URL https://aclanthology.org/2021.tacl-1.13/.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

European Union. General data protection regulation (gdpr), article 22: Automated individual decision-making, including profiling. Regulation (EU) 2016/679, 2016. URL https://gdpr.eu/article-22-automated-individual-decision-making-including-profiling/. Accessed: 2025-12-28.

Fang, J., Jiang, H., Wang, K., Ma, Y., Shi, J., Wang, X., He, X., and Chua, T.-S. Alphaedit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HvSytvg3Jh.

Ferreira, P., Aziz, W., and Titov, I. Truthful or fabricated? using causal attribution to mitigate reward hacking in explanations. *arXiv preprint arXiv:2504.05294*, 2025.

Fiotto-Kaufman, J. F., Loftus, A. R., Todd, E., Brinkmann, J., Pal, K., Troitskii, D., Ripa, M., Belfki, A., Rager, C., Juang, C., Mueller, A., Marks, S., Sharma, A. S., Lucchetti, F., Prakash, N., Brodley, C. E., Guha, A., Bell, J., Wallace, B. C., and Bau, D. Nnsight and NDIF: democratizing access to open-weight foundation model internals. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=MxbEiFRf39.

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., and Bau, D. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. arXiv:2308.14761.

Gao, L. An ambitious vision for interpretability, December 2025. URL https://www.alignmentforum.org/posts/Hy6PX43HGgmfiTaKu/an-ambitious-vision-for-interpretability. Alignment Forum.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health*, 3 (11):e745–e750, 2021.

Gorton, L., Wang, N., Nguyen, N., Deng, M., Ho, E., Balsam, D., and McGrath, T. Interpreting evo 2: Arc institute's next-generation genomic foundation model. *Goodfire Research*, February 2025. URL https://www.goodfire.ai/research/interpreting-evo-2.

Gottesman, D. and Geva, M. Estimating knowledge in large language models without generating a single token. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3994–4019, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-

main.232. URL https://aclanthology.org/2024.emnlp-main.232/.

Greenblatt, R., Nanda, N., Buck, and habryka. How useful is mechanistic interpretability?, 2023. URL https://www.lesswrong.com/posts/tEPHGZAb63dfq2v8n/how-useful-is-mechanistic-interpretability. LessWrong Blogpost.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Gur-Arieh, Y., Suslik, C. H., Hong, Y., Barez, F., and Geva, M. Precise in-parameter concept erasure in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18997–19017, 2025.

Haklay, T., Orgad, H., Bau, D., Mueller, A., and Belinkov, Y. Position-aware automatic circuit discovery. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2792–2817, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.141. URL https://aclanthology.org/2025.acl-long.141/.

Haklay, T., Orgad, H., Reusch, A., Mosbach, M., Wiegreffe, S., Tenney, I., and Geva, M. 1st Actionable Interpretability Workshop at ICML, July 2025b. URL https://icml.cc/virtual/2025/workshop/39962.

Han, X., Wallace, B. C., and Tsvetkov, Y. Explaining black box predictions and unveiling data artifacts through influence functions. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5553–5563, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.492. URL https://aclanthology.org/2020.acl-main.492/.

Havaldar, S., Jin, H., Kim, C., Xue, A., You, W., Gatti, M., Jain, B., Qu, H., Hashimoto, D. A., Madani, A., et al. T-fix: Text-based explanations with features interpretable to experts. *arXiv preprint arXiv:2511.04070*, 2025.

Hendrycks, D. and Hiscott, L. The misguided quest for mechanistic ai interpretability. *AI Frontiers*, May 2025. URL https://ai-frontiers.org/articles/the-misguided-quest-for-mechanistic-ai-interpretability.

Ho, E. On optimism for interpretability, July 2025. URL https://www.goodfire.ai/blog/on-optimism-for-interpretability. Goodfire AI Blog.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Huang, J., Tao, J., Icard, T., Yang, D., and Potts, C. Internal causal mechanisms robustly predict language model out-of-distribution behaviors, 2025. URL https://arxiv.org/abs/2505.11770.

Jacovi, A. and Goldberg, Y. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310, 2021.

Jain, S., Wiegreffe, S., Pinter, Y., and Wallace, B. C. Learning to faithfully rationalize by construction. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Jin, H., Havaldar, S., Kim, C., Xue, A., You, W., Qu, H., Gatti, M., Hashimoto, D. A., Jain, B., Madani, A., et al. The fix benchmark: Extracting features interpretable to experts. *arXiv preprint arXiv:2409.13684*, 2024.

Jin, H., Xue, A., You, W., Goel, S., and Wong, E. Probabilistic stability guarantees for feature attributions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Karimi, A.-H., Schölkopf, B., and Valera, I. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362, 2021.

Kim, C., You, W., Havaldar, S., and Wong, E. Evaluating groups of features via consistency, contiguity, and stability. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. In *Explainable AI:*

Interpreting, explaining and visualizing deep learning, pp. 267–280. Springer, 2019.

King, F., Pettersen, C., Posselt, D., Ringerud, S., and Xie, Y. Leveraging sparse autoencoders to reveal interpretable features in geophysical models. *Journal of Geophysical Research: Machine Learning and Computation*, 2(4): e2025JH000769, 2025.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, July 2020a.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020b. URL https://arxiv.org/abs/2007.04612.

Krishnan, M. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3):487–502, 2020.

Lai, S., Hu, L., Wang, J., Berti-Equille, L., and Wang, D. Faithful vision-language interpretation via concept bottleneck models. In *The Twelfth International Conference on Learning Representations*, 2024.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Li, V. R., Kaufmann, J., Wattenberg, M., Alvarez-Melis, D., and Saphra, N. Can interpretation predict behavior on unseen data?, 2025. URL https://arxiv.org/abs/2507.06445.

Li, Z., Ji, J., and Zhang, Y. From kepler to newton: Explainable ai for science discovery. In *ICML 2022 2nd AI for Science Workshop*.

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

Liu, K., Casper, S., Hadfield-Menell, D., and Andreas, J. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.291. URL https://aclanthology.org/2023.emnlp-main.291/.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning. In Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., and Krisnadhi, A. A. (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.20.

Ma, C., Donnelly, J., Liu, W., Vosoughi, S., Rudin, C., and Chen, C. Interpretable image classification with adaptive prototype-based vision transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Magnusson, I., Tai, N., Bogin, B., Heineman, D., Hwang, J. D., Soldaini, L., Bhagia, A., Liu, J., Groeneveld, D., Tafjord, O., Smith, N. A., Koh, P. W., and Dodge, J. Datadecide: How to predict best pretraining data with small experiments, 2025. URL https://arxiv.org/abs/2504.11393.

Mangla, P., Singh, V., and Balasubramanian, V. N. On saliency maps and adversarial robustness. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*, pp. 272–288, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-67660-5. doi: 10.1007/978-3-030-67661-2_17. URL https://doi.org/10.1007/978-3-030-67661-2_17.

Marks, S. Principles for picking practical interpretability projects, July 2025. URL https://www.lesswrong.com/posts/DqaoPNqhQhwBFqWue/principles-for-picking-practical-interpretability-projects. Alignment Forum.

Marks, S., Hase, P., and Alignment Science Team. Recommendations for technical ai safety research directions, January 2025. URL https://alignment.anthropic.com/2025/

recommended-directions/. Anthropic Alignment Science Blog.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., et al. Artificial intelligence index report 2025. In *Artificial Intelligence Index Report 2025*. 2025.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Meng, K., Sharma, A. S., Andonian, A. J., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

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

Mosbach, M., Gautam, V., Vergara Browne, T., Klakow, D., and Geva, M. From insights to actions: The impact of interpretability and analysis research on NLP. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3078–3105, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.181. URL https://aclanthology.org/2024.emnlp-main.181/.

Mueller, A., Geiger, A., Wiegreffe, S., Arad, D., Arcuschin, I., Belfki, A., Chan, Y. S., Fiotto-Kaufman, J. F., Haklay, T., Hanna, M., Huang, J., Gupta, R., Nikankin, Y., Orgad, H., Prakash, N., Reusch, A., Sankaranarayanan, A., Shao, S., Stolfo, A., Tutek, M., Zur, A., Bau, D., and Belinkov, Y. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=sSrOwve6vb.

Nanda, N. A longlist of theories of impact for interpretability, 2022. URL https://www.alignmentforum.org/posts/uK6sQCNMw8WKzJeCQ/a-longlist-of-theories-of-impact-for-interpretability. LessWrong Blogpost. Accessed: 2025-11-24.

Nanda, N. and Bloom, J. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Nanda, N., Engels, J., Conmy, A., Rajamanoharan, S., bilalchughtai, McDougall, C., Kramár, J., and Smith, L. A pragmatic vision for interpretability. https://www.alignmentforum.org/posts/StENzDcD3kpfGJssR/a-pragmatic-vision-for-interpretability, 2025. Accessed: 2025-12-02.

Newman, B., Ravichander, A., Jung, J., Xin, R., Ivison, H., Kuznetsov, Y., Koh, P. W., and Choi, Y. The curious case of factuality finetuning: Models' internal beliefs can improve factuality, 2025. URL https://arxiv.org/abs/2507.08371.

Obeso, O., Arditi, A., Ferrando, J., Freeman, J., Holmes, C., and Nanda, N. Real-time detection of hallucinated entities in long-form generation, 2025. URL https://arxiv.org/abs/2509.03531.

Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Olah, C. Interpretability dreams. Transformer Circuits Thread Blogpost, May 2023. URL https://transformer-circuits.pub/2023/interpretability-dreams/index.html.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Orgad, H., Kawar, B., and Belinkov, Y. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Peysakhovich, A. and Lerer, A. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.

Potts, C. Assessing skeptical views of interpretability research. Blogpost, July 2025. URL https://web.stanford.edu/~cgpotts/blog/interp/.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

Prenosil, G. A., Weitzel, T. K., Bello, S. C., Mingels, C., Manzini, G., Meier, L. P., Shi, K.-Y., Rominger, A., and Afshar-Oromieh, A. Neuro-symbolic ai for auditable cognitive information extraction from medical reports. *Communications Medicine*, 5(1):491, 2025.

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.647/.

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

Saranti, A., Hudec, M., Mináriková, E., Takáč, Z., Großschedl, U., Koch, C., Pfeifer, B., Angerschmid, A., and Holzinger, A. Actionable explainable ai (axai): a practical example with aggregation functions for adaptive classification and textual explanations for interpretable machine learning. *Machine Learning and Knowledge Extraction*, 4(4):924–953, 2022.

Schut, L., Tomašev, N., McGrath, T., Hassabis, D., Paquet, U., and Kim, B. Bridging the human–ai knowledge gap through concept discovery and transfer in alphazero. *Proceedings of the National Academy of Sciences*, 122(13):e2406675122, 2025. doi: 10.1073/pnas.2406675122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2406675122.

Segerie, C.-R. Against almost every theory of impact of interpretability, 2023. URL https://www.lesswrong.com/posts/LNA8mubrByG7SFacm/against-almost-every-theory-of-impact-of-interpretability-1. LessWrong Blogpost. Accessed: 2024-08-17.

Shah, R., Irpan, A., Turner, A. M., Wang, A., Conmy, A., Lindner, D., Brown-Cohen, J., Ho, L., Nanda, N., Popa, R. A., et al. An approach to technical agi safety and security, 2025. URL https://arxiv.org/abs/2504.01849.

Shovelain, J. and Mckernon, E. The risk-reward tradeoff of interpretability research, July 2023. URL https://www.lesswrong.com/posts/HdqdqNC3MyABHzSqf/the-risk-reward-tradeoff-of-interpretability-research. LessWrong.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Singh, R., Miller, T., Lyons, H., Sonenberg, L., Velloso, E., Vetere, F., Howe, P., and Dourish, P. Directive explanations for actionable explainability in machine learning applications. *ACM Trans. Interact. Intell. Syst.*, 13(4), December 2023. ISSN 2160-6455. doi: 10.1145/3579363. URL https://doi.org/10.1145/3579363.

Singh, R., Miller, T., Sonenberg, L., Velloso, E., Vetere, F., Howe, P., and Dourish, P. An actionability assessment tool for explainable ai. *arXiv preprint arXiv:2407.09516*, 2024.

So8res. If interpretability research goes well, it may get dangerous, April 2023. URL https://www.lesswrong.com/posts/BinkknLBYxskMXuME/if-interpretability-research-goes-well-it-may-get-dangerous. LessWrong.

Spillner, L., Ringe, R., Porzel, R., and Malaka, R. Can ai explanations make you change your mind? *arXiv preprint arXiv:2508.08158*, 2025.

Steinhardt, J. and Schwettmann, S. Introducing transluce, October 2024. URL https://transluce.org/introducing-transluce. Transluce AI Blog.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Upadhyay, S. and Barez, F. Martian interpretability challenge, part 2: The core problems in interpretability. Martian Blog, December 2025. URL https://withmartian.com/post/interpretability-prize-part2. Accessed: 2025-12-08.

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 10–19. ACM, 2019.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C. K., Suter, P., and Thijs, L. G. The sofa (sepsis-related organ failure

assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix). *Intensive care medicine*, 22(7):707–710, 1996.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Wen, X., Tan, W., and Weber, R. O. Gaprotonet: A multi-head graph attention-based prototypical network for interpretable text classification, 2024.

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., and Potts, C. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962, 2024.

Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., and Potts, C. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

Xue, A., Alur, R., and Wong, E. Stability guarantees for feature attributions with multiplicative smoothing. *Advances in Neural Information Processing Systems*, 36: 62388–62413, 2023.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023.

Yang, Y., Gandhi, M., Wang, Y., Wu, Y., Yao, M. S., Callison-Burch, C., Gee, J. C., and Yatskar, M. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 90683–90713. Curran Associates, Inc., 2024.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

You, W., Qu, H., Gatti, M., Jain, B., and Wong, E. Sum-of-parts: Self-attributing neural networks with end-to-end learning of feature groups. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=r6y9TEdLMh.

You, W., Xue, A., Havaldar, S., Rao, D., Jin, H., Callison-Burch, C., and Wong, E. Probabilistic soundness guarantees in llm reasoning chains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7517–7536, 2025b.

Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Zhao, X., Yin, F., and Durrett, G. Understanding synthetic context extension via retrieval heads. In *Forty-second International Conference on Machine Learning*, 2025.

Zhou, Z., Lin, X., Xu, X., Prakash, A., Rus, D., and Low, B. K. H. DETAIL: Task DEmonstration attribution for interpretable in-context learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4jRNkAH15k.

A. Visualizing the space of Actionable Interpetability work

Figure 3 demonstrates the different types of actionable interpretability work as spanned by the dimensions of actionability.

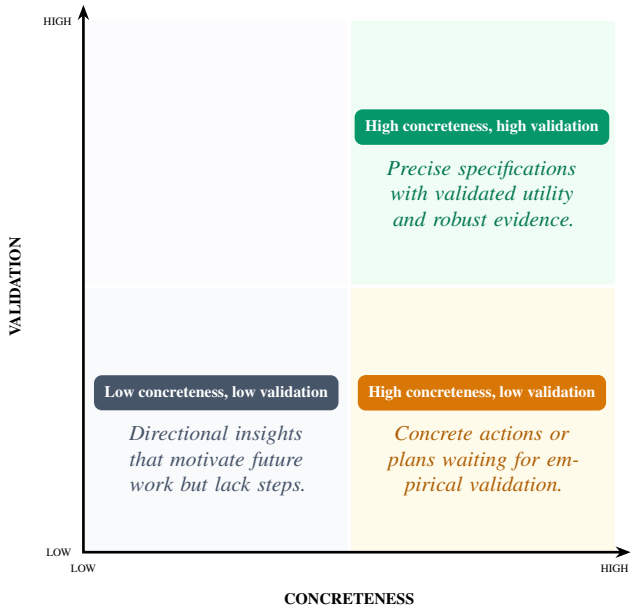

Figure 3. The Space of Actionable Interpretability Work.

B. Additional Examples of Actionable Work

B.1. Actions That Modify Model's Output

This appendix provides additional examples of actionable interpretability work that modifies model's output.

Data curation. Pruthi et al. (2020) estimate the training data influence by tracing how the loss on a test example changes due to each training example during gradient descent. Mangla et al. (2020) use saliency maps to guide adversarial training. They demonstrated that they can identify mislabeled training examples and data poisoning attacks, and that removing the most negatively influential training examples improves model performance. Magnusson et al. (2025) show that strategic data filtering enables small-scale evaluations to forecast large-scale benchmark performance at orders-of-magnitude lower compute cost.

Model input. Peysakhovich & Lerer (2023) used attention analysis to discover that models attend more to relevant documents even when not using them, then developed "attention sorting"—reordering documents at inference time to improve retrieval-augmented generation performance.

Training decisions. Newman et al. (2025) build on findings that models' internal representations of truthfulness can contradict their outputs (Liu et al., 2023; Orgad et al., 2025). They use these internal representations to curate post-training data, selecting the model's own generations that align with its internal signals of truthfulness, and demonstrate that this approach reduces hallucinations.

Self-explaining models. Models that generate explanations as an integral part of their prediction process offer a distinct form of actionability: users can inspect and potentially intervene on the intermediate reasoning. This paradigm includes self-attribution architectures (Agarwal et al., 2021; Brendel & Bethge, 2019; Jain et al., 2020), interpretability wrappers for foundation models (You et al., 2025a), selecting prototypes (Ma et al., 2024; Wen et al., 2024), predicting directly off of concepts (Koh et al., 2020a; Yang et al., 2023; Lai et al., 2024; Yang et al., 2024), or generating programs as explanations that calculate the outcome (Lyu et al., 2023).

B.2. Actions About Deployment and Use

End user decisions. Recent work demonstrates how the internal representations of LLMs can provide uncertainty estimations about their outputs, enabling users to detect errors and make informed decisions (Kadavath et al., 2022; Azaria & Mitchell, 2023; Gottesman & Geva, 2024; Orgad et al., 2025, inter alia). Building on this foundation, Chakraborti et al. (2025) argue that AI systems in high-stakes domains like healthcare must provide personalized uncertainty estimates to support decision-making: when a clinical decision support system indicates high uncertainty, clinicians can choose to override recommendations rather than following potentially unreliable outputs. Obeso et al. (2025) presented a method for real-time identification of hallucinated tokens in long-form generations.

Chen et al. (2024b) developed a dashboard that exposes the model's internal "user model"" in real time; user studies showed participants valued this transparency for identifying biased behavior.

Deployment decisions. Although Chen et al. (2024a) use a separate scoring function rather than interpretability techniques directly, it illustrates how uncertainty estimates enable benefits. Casper et al. (2024b) organized a competition to evaluate whether interpretability tools could help humans detect backdoors implanted in ImageNet-scale CNNs using feature synthesis methods inspired by interpretability research. They achieved 49% human detection rates, significantly outperforming dataset-based attribution methods.

B.3. Shaping Future Practice

Learning from superhuman models. Goodfire's interpretation of Arc Institute's biological foundation model Evo 2 (Gorton et al., 2025) identifies biologically relevant structure in model representations, demonstrating interpretability's potential to guide scientific investigation.

C. Additional Examples on Evaluating Actionability

C.1. Evaluating Understandability

In high-stake decisions contexts, interpretability must present model behavior in a form that aligns with users' existing conceptual frameworks in order to be acted upon. Understandability can be evaluated by measuring how well explanations align with a user's domain-specific reasoning. Benchmarks such as Features Interpretable to eXperts (FIX) (Jin et al., 2024) and its textual extension T-FIX (Havaldar et al., 2025) operationalize this notion by assessing whether explanations correspond to established concepts in domains such as astrophysics or medicine (e.g., cosmological structures or clinical scoring systems like SOFA (Vincent et al., 1996)). For non-expert users, understandability is often captured through plausibility metrics (Agarwal et al., 2024), which evaluate whether explanations appear coherent and reasonable given common-sense expectations.

C.2. Evaluating Reliability

A large body of prior work proposes metrics for evaluating the robustness of explanations, including sensitivity of feature attributions (Alvarez-Melis & Jaakkola, 2018; Yeh et al., 2019; Kindermans et al., 2019), explanation invariance (Crabbé & van der Schaar, 2023), and provable guarantees on explanation behavior (Blanc et al., 2021; Bassan & Katz, 2023).

One example is work on robustness guarantees in the form of *stability certificates* (Xue et al., 2023; Kim et al., 2024). These certificates explicitly quantify how sensitive a model's predictions are to changes implied by an explanation, such as removing or altering explanatory features. More recent work has extended such guarantees to large-scale foundation models (Jin et al., 2025), chain-of-thought explanations (You et al., 2025b), and clinical applications such as Alzheimer's disease (Achara et al., 2025).