Not All Interpretation Methods Are Created Equal: Tiered Evaluation of AI Interpretability

NICOLAS SANCHEZ and BIN YU✉, University of California, Berkeley, USA

As AI transitions to sociotechnical deployments, uninterpretable models pose serious threats to reliability, public trust and contestability. However, not all interpretation methods are created equal: for interpretability to enhance deployment safety and effectiveness, interpretation methods must be evaluated in a tiered system to differentiate evidence levels of trustworthiness, rather than validated through isolated interpretation instances.

In this paper, we advocate elevating the evaluation of interpretation methods as a core assessment metric for AI system deployment. To this end, we propose TRUST, a tiered evaluation framework that builds upon the Predictability-Computability-Stability (PCS) principles for Veridical Data Science (VDS) [22, 30, 31, 33] and the Predictive, Descriptive and Relevance (PDR) [19] foundations. To differentiate interpretation methods effectively and systematically, TRUST grades interpretation across five tiers: Task-Aligned PDR Foundations, Rigorous Epistemic Grounding, Unifying Semantic Understanding, Scope Transferability, and Targeted Prospective Validation. We present a start-to-finish case study of the iterative Random Forest (iRF) family under TRUST and contrast it with other LLM interpretation methods that struggle to clear specific tiers. Finally, we conclude with brief initial thoughts on how TRUST principles can extend in autonomous agentic frameworks. Ultimately, the TRUST framework fosters contestability and epistemic grounding as foundational evaluation principles for human-machine workflows and, eventually, autonomous agentic workflows.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Artificial Intelligence**; • **Human-centered computing** → **Human Computer Interaction (HCI)**.

Additional Key Words and Phrases: Interpretable Machine Learning, Evaluation Frameworks, TRUST Framework, Veridical Data Science, Contestability, Predictability-Computability-Stability (PCS), AI Governance

ACM Reference Format:
Nicolas Sanchez and Bin Yu. 2026. Not All Interpretation Methods Are Created Equal: Tiered Evaluation of AI Interpretability. 37, 4, Article 111 (October 2026), 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Artificial Intelligence (AI) models are rapidly transitioning from closed-loop novelties into embedded sociotechnical systems. From NHS England providing over half a million clinicians with AI tools [21], to US Census data showing 37% of large firms and 55% of workers using AI [20, 26], to expanding classroom partnerships [13], AI is deeply integrated into daily workflows. However, rapid adoption has triggered unintended yet severe vulnerabilities driven by user complacency and automation bias: clinicians override their own accurate medical judgment [24], lawyers unwittingly submit fabricated precedents [25], and educators wrongly accuse non-native speaking students of misconduct [17]. These breakdowns have sparked significant reactions, including Norway banning AI from elementary schools [2], California establishing state AI safety verification frameworks [10], and congressional investigations into UnitedHealth's nH Predict for improperly denying

Authors' Contact Information: Nicolas Sanchez, nicolas_sanchez_ep@berkeley.edu; Bin Yu, binyu@berkeley.edu, University of California, Berkeley, Berkeley, CA, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM XXXX-XXXX/2026/10-ART111
https://doi.org/XXXXXXX.XXXXXXX

care [8]. Ultimately, these examples reveal a systemic flaw: current evaluation frameworks rely on out-of-context benchmarks that ignore deployment realities. While paradigms like Green Shielding [16] evaluate outputs under everyday variations and propose mitigation strategies to address failure modes, AI systems must go further by satisfying the foundational societal principle of contestability [27].

Contestability has historically referred to mechanisms for users to understand, construct, shape and challenge model predictions or decisions [18]. These mechanisms establish accountability and empower steerability in socially deployed systems. Bamberger, Mulligan in [4], and Wang have demonstrated the need to expand this definition beyond an ex-post feature: contestability must be a systemic property. AI systems must possess reflexive architectures, designed to surface embedded value choices and provide structured pathways to challenge model reasoning not just at the output level, but at configuration and design time. Trustworthy model interpretability offers an indispensable avenue for addressing these requirements in addition to guiding trustworthy model development. However, not all interpretation methods are created equal. Fragile, unreliable interpretation methods used to address contestability are counter-productive: by providing erroneous perceptions of model reasoning the false sense of security deepens societal mistrust in AI systems.

To better evaluate trustworthy interpretation methods for trustworthy deployment and contestability requirements, we propose a five-tiered evaluation framework designed to grade interpretation methods for fidelity, safety and utility. Each tier demands progressively stronger independent evidence starting from baseline requirements to gold-standard external validation: Task-Aligned Predictive, Descriptive, and Relevant (PDR) Foundations; Rigorous Epistemic Grounding; Unifying Semantic Understanding; Scope Transferability; and Targeted Prospective Validation (TRUST). In conjunction with the proposed framework, we also provide a standardized TRUST Card [1] designed both for documented evaluation and guided development. The TRUST tiers build upon two established pillars towards trustworthy and safe AI including contestability. First, it directly expands the foundational definitions of interpretable machine learning set forth by Murdoch et al. [19]. Second, it relies on the principles of Veridical Data Science (VDS) [22, 30, 31, 33].

Within Veridical Data Science (VDS), the core principles of Predictability, Computability, and Stability (PCS) examine reasonable human judgment calls and aggregate good or reality-checked ones across every stage of the AI pipeline [33]. Grounded in PCS, the TRUST framework assesses the trustworthiness of interpretation methods throughout the AI lifecycle—from initial design, training and pre-deployment validation to live monitoring and post-mortem auditing. To accommodate diverse stakeholders with varying domain and technical expertise, TRUST provides a multi-dimensional, tiered evaluation framework where later tiers demand progressively stronger external independent evidence.

For the remainder of this paper, we detail the five tiers of TRUST, followed by a start-to-finish demonstration using the iterative Random Forest (iRF) in biological settings as well as other LLM interpretation methods (CD-T, SPEX, GCT). In our concluding discussion, we share initial thoughts on extending TRUST to autonomous multi-agent systems.

2 Bottom-Up TRUST Evaluation Framework

2.1 Tier 1 : Task-Aligned Predictive, Descriptive, and Relevant (PDR) Foundations

"There's no sense in being precise when you don't even know what you're talking about."
— John von Neumann (Method in the Physical Sciences)

The first and most basic tier directly builds on the formalization of interpretability introduced by

[1] An interactive form with the TRUST card is available at https://forms.gle/jro7XjNhWvi6wDD18

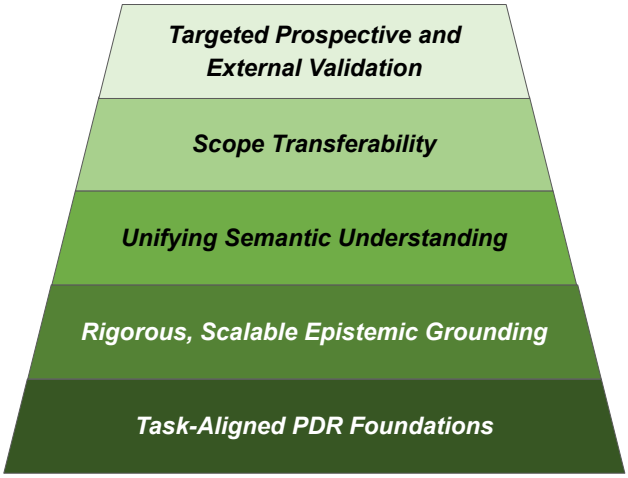


Fig. 1. Bottom-Up TRUST Evaluation Framework

Murdoch et al. [19]. An interpretation method, and its associated AI system must, as a baseline, satisfy three desiderata to ensure it provides both reliable and sociotechnically actionable insights:

Reality Check ("P"). The scope of predictive accuracy broadens to encompass the strength of the AI system's reality checks, ensuring that the interpreted system captures true underlying relationships in the data. Notably, accuracy should not be restricted to standard average accuracy. Instead, evaluation requires multiple metrics for accuracy (e.g. sensitivity, specificity, worst-case error, top-K) stratified across key subpopulations that are tailored for the task.

Descriptive Accuracy ("D"). An interpretation method's outputs must faithfully reflect the AI system's internal decision-making logic. This includes the need for outputs to be cleanly translatable and understandable to the desired stakeholders without sacrificing fidelity.

Relevance ("R"). This demands that the communicated information proves useful to specific target stakeholders for a specific task. Crucially, the choice of task and target stakeholders is directly tied to the purpose assigned to the AI system being interpreted, and the human workflow it is to be embedded within. This is especially important for contestability: a descriptive explanation that fails to incorporate normative foundations becomes irrelevant for contestation since it falls short of empowering subjects to meaningfully challenge or affect the underlying value alignment of the system.

2.2 Tier 2 : Rigorous and Scalable Epistemic Grounding

"[T]he user of anything is the one who knows most of it by experience[...] . As, for example, the flute-player reports to the flute-maker which flutes respond and serve rightly in flute-playing, and will order the kind that must be made, and the other will obey and serve him. "

— Plato's Republic Book X

Trusted incumbent domain experts provide a uniquely critical reality check ("P" in PCS). They both possess unique contextual knowledge of domain-specific dynamics and are responsible for ensuring that the AI system does not corrupt the workflow's purpose upon integration [18]. However, with the growth of data availability, AI outputs have scaled with more compute at a much faster pace than human expert bandwidth. Exhaustive, instance-by-instance review thus becomes unviable. This

scaling bottleneck directly connects to the computability principle ("C" in PCS) that here extends from algorithmic computation to encompass resource-realistic feasibility in terms of human time and cognitive effort. Tier 2, Rigorous and Scalable Epistemic Grounding, demands that interpretation methods explicitly identify this trustworthy expert and structure their outputs to enable systematic, efficient review at scale. Three criteria are of notable importance for addressing Tier 2.

Objective Expert Validation. The expert or other reference that will serve to review the interpretation method outputs must be independent, established and trusted sources of domain knowledge. The presentation of the output must actively mitigate confirmation bias and automation complacency during review. Interpretation methods often risk being persuasive by presenting plausible-seeming rationalizations for system decisions. To fight this, review procedures should emphasize controlled methodologies, including blinded expert evaluations and presentation of negative controls [1, 9].

Cognitively Compatible Representations. The interpretation method output presentation should strive to be compatible with established domain terminology or format (e.g. mathematical equations, logical statements or standardized rubrics). By presenting outputs in a familiar, easy-to-reference format, limited expert cognitive resources can focus on critical decision logic rather than parsing novel presentation. However, this push for compatibility should not come at the cost of precluding an interpretation method from presenting a novel form of insight that challenges existing paradigms.

Scalable Auditing Process. The selection of data or examples for expert review must intentionally balance realistic expert bandwidth with rigor, directly aligning with the sociotechnical purpose and priorities of the system (e.g. outputs close to the model decision boundary or within high-risk subpopulations). Furthermore, clustering techniques can group recurring logic and enable structural rather than isolated instance level review. When tested for safety and trustworthiness, specialized expert AI systems can also be integrated to perform semi-automated auditing.

2.3 Tier 3 : Unifying Semantic Understanding

"If names are not correct, language will not be in accordance with the truth of things. "
— Confucius (The Analects)

In line with the stability principle ("S") of PCS, Tier 3, Unifying Semantic Understanding among Stakeholders, evaluates whether an interpretation method's outputs provide unambiguous and stable insights across relevant stakeholders.

Baseline Output Stability. Fundamentally, unified understanding requires stability and consistency in the interpretation method's outputs against reasonable perturbations within the context of the task. Similar inputs processed through the same internal model (within reasonable data perturbations or random noise) must yield functionally similar interpretation outputs. Otherwise, a fragile interpretation method will inevitably lead to conflicting conclusions.

Linguistic Stability. Jargon collision between disciplines is of particular concern. The word "matrix" may refer to tabular data for an engineer but relate to cellular micro-environments for a biologist. Similarly, "bias" may mean a systematic deviation of an estimator for a statistician versus suggesting issues of fairness and discrimination to a regulator. Interpretation methods must communicate the underlying decision-process using shared and stable language that reduces the likelihood of conflicting assumptions.

Representative Examples. The unified semantic understanding should be validated across intentionally chosen and clearly documented representative examples. Following the Population, Question of interest, Representativeness and Scrutiny (PQRS) framework [32], representativeness assesses how closely the chosen examples relate to the target population of potential deployment-usages . The statistical definition of population is the complete set of conditions, environments, and subpopulations under which the data and the output are generated in the real-world [32]. Notably, the usage of the interpretation method may differ slightly from the target population of the underlying deployed AI system. Interpretation is often disproportionately invoked in high-consequence or edge-case decisions where agreement from different stakeholders is most critical. Formal sampling strategies such as stratified random sampling across domain-critical subgroups, or decision boundary regions should be carefully documented (e.g. sampling method description or code, inclusion/exclusion criteria) to convincingly justify that these chosen examples achieve target population representativeness. If random sampling is not used, as often is the case for benchmarks, documentation is needed to make the case to describe the target population for the benchmark and how it achieves representativeness.

2.4 Tier 4 : Scope Transferability

"To foresee consists in seeking to apply to a new case what we have observed in a former case; but the new case will never be identical with the former."

— Henri Poincaré (Science and Hypothesis)

The stability principle ("S") also requires that interpretation methods intended for real-world deployment remain robust against new sources of datasets and related or even far-away domains. At deployment, the AI system will be exposed to a variety of deployment conditions and, as such, distinct target populations. Data curators may use different labeling standards or record data using different hardware. Engineers may store and pre-process that data through different pipelines. Domain practitioners may employ differing protocols for how and when the AI system is used. As a result, it is critical to understand if the interpretation methods remain trustworthy under these changes. Tier 4, Scope Transferability, therefore evaluates an interpretation method's proven capacity to generalize beyond its original dataset and context, to a broader class of datasets or related or even far-away tasks.

Broad Validation. Historically, supervised learning models have been rigorously vetted against at least dozens of benchmark datasets if not hundreds to prove their efficacy. In contrast, interpretability methods are frequently validated on only a handful of highly specific examples. Tier 4 instead demands empirical validation across a broad spectrum of public and private datasets to match the scale of early supervised learning by testing across at least dozens of diverse tasks.

Demarcation of Applicability. No single interpretation method will work for all possible domains. Nevertheless, testing across a variety of datasets provides evidence to clearly define the areas of applicability of the method.

Controlled Setting Validation. One notable form of scope transferability covered under Tier 4 involves testing interpretation methods under controlled conditions with well-understood ground truth. Demonstrating the success of methods in recovering those ground-truth dynamics in realistic synthetic settings or simplified but still realistic versions of the task provides evidence of the interpretation method's performance before justifying its applicability in more open-ended, real world environments.

2.5 Tier 5 : Targeted Prospective and External Validation

"Experience does not err. Only your judgments err by expecting from her what is not in her power. "
— Leonardo Da Vinci (Codex Atlanticus)

Tier 5, Targeted Prospective and External Validation, demands what is, at the moment, arguably the ultimate end-goal of interpretation methods: to generate insights that lead stakeholders to discover new, actionable knowledge verifiable through targeted prospective and external validation. While Tier 4 touches on this by focusing on robustness on retrospective datasets, Tier 5 goes further by going from observational evaluation to prospective real-world execution.

To clear Tier 5, interpretation method outputs must surface concrete testable targets for prospective intervention. This could include novel targeted data collection, prospective wet-lab experiments or randomized intervention field trials. Tier 5 further demands that these actions result in empirically confirmed positive outcomes, defined in close relation to the purpose of the underlying AI system and the relevant stakeholders. For scientific discovery, an interpretation method may lead to a high-replicability rate experiment that justifies a novel scientific law or principle. For healthcare or public policy, the interpretation method may lead to intervention that improves health outcomes or quality of life. The interpretation method should make clear what the positive outcome is and who it benefits with clear documentation.

The transition from pure algorithmic insight to real-world experimentation carries a strong implicit endorsement for deployment: to suggest and execute a relevant physical experiment requires a deeper understanding of operational dynamics and feasibility. Clearing Tier 5 establishes the successful execution of intentional, external and prospective validation as the strongest proof of an interpretation's actionable relevance and real-world stability within a purpose-driven system.

3 A Gold Standard in Biology: Iterative Random Forest (iRF)

The TRUST framework is broadly applicable across a wide class of AI systems. However, the core principles were inspired by the development of iterative Random Forest (iRF) [5, 15, 28], an algorithm (and its variants) that adds stability to random forests to enable interpretable interactions. Evaluating iRF serves as a start-to-finish demonstration of TRUST.

iRF operates under a binary classification setup where stability is added to random forests (RF) through weighted sampling of the features for splits based on MDI importance scores from the last iteration and the first iteration uses RF. Two features that fall on the same path of a tree in the forest are interpreted as interacting. Random Intersection Trees (RIT) traverse the paths and extract candidate Boolean feature interactions. Interactions are assigned stability scores via outer-layer bootstrapping, and only stable ones are preserved.

The base iRF identifies importances from single features and Boolean feature interactions. Signed iRF (siRF) incorporates the directional nature of the feature splits while low-signal iRF (lo-siRF) adapts the pipeline to low-signal settings. We consider iRF across three biological settings:

Drosophila Embryogenesis iRF predicts enhancer status based on transcription factor bindings, histone modifications and chromatin accessibility to guide developmental biologists [5, 15]

Alternative Splicing in Human Cells iRF predicts RNA splicing rates from histone modifications and binding proteins to help biologists uncover co-transcriptional regulatory mechanisms [5]

Hypertrophic Cardiomyopathy (HCM) iRF predicts left ventricular mass (LVM) to help cardiovascular biologists discover gene and multi-gene epistatic interactions responsible for a specific heart disease (HCM) [28].

3.1 Tier 1: Task-Aligned PDR Foundations

iRF satisfies the PDR Foundations across all three biological settings.

"P". In Drosophila, iRF matched state-of-the-art standard Random Forests. In human splicing, iRF increased held-out test set performance over random forests (from 0.44 to 0.51 AUC-PR) while concentrating tree decision paths on fewer features. In HCM, a lower signal regime due to the complex human disease, lo-siRF, slightly surpassed other baselines in classification accuracy while utilizing Boolean interactions much closer to biologically backed epistatic interactions [28].

"D" The outer-loop bootstrap mechanism identifies stable Boolean decision paths that directly reflect the internal mechanics of the predictive engine.

"R" Across all three studies, predicted outcomes (enhancer activity, splicing rates, LVM phenotype) are relevant biological targets. Similarly, inputs and interpretation method outputs (epigenetic markers, single-nucleotide variants) align with accepted biological mechanical drivers of each outcome. Decision tree thresholding also more faithfully reflects biological expression thresholds in biomolecular interaction processes.

3.2 Tier 2: Rigorous and Scalable Epistemic Grounding

iRF also addresses Tier 2 through intentional interactive review by experts or by comparing against established domain knowledge.

Rigor In Drosophila, 20 second-order interactions uncovered by iRF were compared against established literature. In HCM study, uncovered gene and gene-gene interactions were reviewed by cardiovascular biologists in a blinded manner with negative controls (mid-ranked interactions and random interactions).

Scalability RIT+bootstrap displays algorithmic scalability, successfully navigating massive search space ($> 10^9$ candidates for HCM). Human review is facilitated by outputting ranked, high-stable Boolean rules that are easily understandable to biologists. However, exceedingly large candidate pools still challenge true scalability of the biologist review.

3.3 Tier 3: Unifying Semantic Understanding (Among Stakeholders)

iRF robustly satisfies Tier 3. iRF and RIT+bootstrap both emphasize high stability features and outputs, ensuring that findings will be consistent across runs and hence reproducible. Further, explicit Boolean logic with domain-meaningful variables makes for clear logic-based findings that provide unambiguous meaning across biologists.

3.4 Tier 4: Scope Transferability

iRF also addresses scope transferability though extension to more datasets and domains are warranted. As demonstrated by the three studies, iRF architecture shows generalization across different biological domains, from fruit flies to human cardiac disease [5, 28]. iRF also displays performance in simplified settings. In Drosophila, the framework is shown to be robust when enhancer activity follows Local Spiky Sparse (LSS) dynamics [6]. lo-siRF also displayed controlled Type 1 error in purely additive genetic simulations and higher statistical power in epistatic simulations (non-linear gene-gene interactions) compared to linear models [23, 28].

3.5 Tier 5: Targeted Prospective and External Validation

iRF finally robustly addresses Tier 5. For Drosophila, 80% of the 20 prioritized order-2 interactions were validated by experiments in the literature, and order-3 and order-4 interactions generated specific, testable hypotheses regarding enhancer activity [5]. For HCM, 80% (4 out of 5) of lo-siRF suggested gene and gene interactions were subsequently confirmed via physical wet-lab experiments [23, 28].

Table 1. TRUST Evaluation Summary Across Case Studies

Method	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
iRF	Robust	Partial	Robust	Partial	Partial
CD-T	Partial	Partial	Partial	Partial	Unaddressed
SPEX	Partial	Partial	Unaddressed	Partial	Unaddressed
GCT	Partial	Partial	Robust	Partial	Robust

4 Other Interpretation Method Case Studies

To further illustrate the TRUST framework, we evaluate additional interpretation methods against relevant tiers. These are not meant as exhaustive evaluations but rather brief demonstrations to other model classes and interpretation paradigms.

4.1 Contextual Decomposition for Transformers (CD-T)

Contextual Decomposition for Transformers (CD-T) is an internal mechanistic interpretation method for Transformers that captures contributions of combinations of input features or internal components. Guided by PCS, CD-T was used to analyze a BERT model fine-tuned on pathology reports to classify primary Gleason scores, a prognostic cancer score [11]. More specifically, CD-T extracts a stable circuit of attention heads from a collection of pathology reports. CD-T then finds influential words by identifying the most highly attended words when averaged across attention maps from that stable circuit. The purpose is to help urologists understand the model reasoning that drives a Gleason score classification. Main stakeholders include the urologist reviewing the BERT prediction seeking clinical rationale, patients whose treatment pathway depends on the cancer score, and healthcare administrators overseeing clinical compliance.

From the perspective of TRUST tiers, CD-T partially addresses Tiers 1 through 3. For Tier 1, the underlying BERT model achieves high accuracy on classification for a well-established prostate cancer score, supporting “R” and "P" in PDR. CD-T also offers higher descriptive faithfulness ("D") than path-patching (Tier 1). CD-T outputs were reviewed by expert urologists, showing text-level attributions highlighted diagnostically relevant criteria for Gleason (Tier 2) and helped them differentiate between more or less performing models. CD-T also supports output stability by being grounded in PCS (Tier 3). However, one semantic gap negatively affects all three tiers: attention-head circuits represent computation flow rather than clinical diagnostic rationale, rendering them ambiguous to non-technical stakeholders. Without clear evidence that clinical experts can consistently map between the two concepts, both the rigor of the expert review and the unified semantic understanding brought about by CD-T are put into question.

For Tier 4, CD-T possesses strong methodological transferability, generalizing across transformer architectures and tested on non-clinical text (e.g. SST-2, AGNews [11, 12]). However, it fails to compel or produce actionable recourse or concrete safety guards based on these outputs, falling short on Tier 5.

4.2 Spectral Explainer (SPEX)

Spectral Explainer (SPEX) focuses on the efficient discovery of model circuits in black box LLMs via scalable Boolean interaction of features [14]. SPEX assumes that interactions are sparse, masks token inputs through random masking design before learning the Fourier transforms using sparse graph decoding tools to find interactions. To reduce the number of LLM calls, ProxySPEX [7] extends SPEX by using a surrogate gradient-boosted tree (GBT) model to take advantage of the hierarchical structure of interactions and achieve a similar level of faithfulness with one tenth of

the expensive LLM calls (emphasizing the "C" in PCS). We consider here the application of SPEX on the HotPotQA dataset [29], a dataset which requires synthesizing facts across different documents to provide a correct answer. The purpose here is to demonstrate the LLM is in fact incorporating and reasoning from these different facts. The stakeholders would be natural language processing (NLP) and AI safety researchers seeking proof of correct LLM reasoning.

Similar to CD-T, SPEX partially addresses Tiers 1-3 in its construction and evaluation but is weakened by a persistent semantic gap. For Tier 1, the evaluated LLMs achieve high benchmark exact match and F1 score on the HotPotQA, ensuring “P” in PDR. Evaluating feature interactions across documents directly addresses the multi-hop reasoning that is being tested ("R"). Algorithmic descriptive accuracy ("D") is verified by achieving better faithfulness (20% higher than marginal independent attribution methods). However, complex boolean interactions between words or phrases remain difficult to interpret for researchers and do not directly align with existing knowledge sources. Indeed, for Tier 2, the epistemic grounding relied on comparing overlaps between SPEX outputs and supporting facts as annotated by humans. Human annotations were collected as a flat set of words, questioning the cognitive compatibility with the boolean interaction outputs of SPEX. This ambiguity combined with lack of interactive review further weakens evidence that these outputs will support a unified semantic understanding among stakeholders (Tier 3).

By virtue of being a model-agnostic, black-box interpretation method, SPEX stands out as highly transferable from a methodological perspective, but is tested only on a handful of datasets. Though SPEX provides audit trails to help detect correct or erroneous reasoning, it stops short of proving actions or actionable hypotheses to affect or improve the model reasoning, falling short on Tiers 4 and 5.

4.3 Generative Causal Testing

PCS-guided Generative Causal Testing [3] (GCT) provides a striking example of integrating Tier 5 into interpretation method design. Within the field of neuroscience, GCT translates complex brain-encoding models into human-readable hypotheses regarding what specific concepts activate certain brain regions.

GCT trains a predictive voxel-wise encoding model mapping LLM embeddings to fMRI BOLD responses (to capture “P” in PDR and PCS). The model is then run across a large corpus, using contrastive prompts and bootstrapping to find the n-grams that consistently elicit the strongest response (“S” in PCS). GCT then outputs an LLM-guided summarization of these n-grams as a candidate hypothesis for the specific voxel or region. Using these outputs as blueprints, an LLM then generates a synthetic, coherent story engineered to trigger the targeted voxel to confirm or refute the hypothesized concepts. Primary stakeholders here include neuroscientists evaluating brain-language selectivity.

Starting with Tier 1, GCT uses fMRI BOLD signals, in line with cognitive neuroscience practice ("R") and initially filters for voxels with high predictive power on held-out test data (“P” from PDR). However, descriptive accuracy, “D” from PDR, is hurt by the LLM inter-mediation. Though a language based summary is clearer to understand for experts (supporting Tiers 2 and 3), the summarized concept can suffer from oversimplification or hallucination by the LLM, severing the direct mechanistic link to the underlying model logic. In terms of epistemic grounding (Tier 2), GCT independently recovered established functional selectivity from previously published neuroscientific literature (such as “Body parts” in the extrastriate body area or “Scenes and settings” in the parahippocampal place area) but lacked an interactive, targeted and blinded review process.

In terms of scope transferability (Tier 4), GCT was shown to work across different scales, from individual voxels to broader cortical regions of interest, including newly found micro regions in the

prefrontal cortex but was not run on a large variety of datasets. Most notably, GCT stands out for its incorporation of Tier 5 in its design. GCT explicitly designs an in-vivo experiment to confirm or refute the output from the interpretation methods. Candidate hypotheses serve as a blue-print to design LLM-assisted synthetic stories structured to validate the hypotheses. The resulting physical brain responses directly confirm or refute the causal validity of the interpretation method output. Success rates of these in vivo experiments varied and were overall not as high as the 80% success rate of iRF.

5 Discussion

We present TRUST as a principled, multi-tiered AI evaluation framework meant both to elevate AI interpretation methods into a standalone core performance measure for AI development, monitoring, and steering as well as to satisfy the foundational principle of contestability for AI governance.

Rigorous evaluation of interpretation methods enables human-machine partnerships. In our view, the ideal integration of AI into workflows mirrors the historical relationship between mathematics and intuition. Humans provide intuition and creative direction grounded by their physical experience. AI systems provide the computational engine to validate, refute or refine those intuitions. Rather than working in parallel, the iteration between human intuition and data-driven AI validation can unlock vast new knowledge, provided trustworthy interpretability serves as the critical bridge. However, elevating interpretability is not intended to restrict AI learning and development to the limits of human cognition. A promise of data-driven AI models is discovering patterns beyond human reach.

To that end, while TRUST's empirical foundations center on human-machine partnerships, the AI landscape compels us to envision expanding TRUST to the emerging paradigm of autonomous multi-agent systems. This prospective adaptation requires a deliberate expansion of core terminology: stakeholders broaden to incorporate machine orchestrators, domain experts will include specialized critic agents. Correspondingly, the evaluation of an interpretation method will extend to the evaluation of inter-agent telemetry, or communication protocols between agents in a multi-agent network. Notably, despite contestability being an inherently human-centric concept, its core mechanisms of transparency, interrogation and steerability are equally critical in this agentic paradigm for machine orchestrators to detect misalignment and prevent silent failure cascades.

Regardless of whether an interpretation method bridges AI to human intuition or to automated AI auditors, when algorithmic systems are deployed in a society they become sociotechnical and demand contestability. AI systems are meant to augment or replace existing entities, whether they be established processes, incumbent domain experts or liable gatekeepers. From a practical point of view, interpretation methods are an important path for the incumbent to vet the soon-to-be-deployed AI system. The tiered TRUST framework becomes indispensable: not all interpretation methods are equally trustworthy for deployment readiness and automated agentic systems require proven telemetry fidelity. Interpretation techniques that satisfy all five tiers of TRUST establish the necessary foundation for deploying, monitoring, and contesting AI systems safely. As the AI deployment landscape continues to evolve, TRUST provides a flexible scaffold that can incorporate additional tiers to maintain alignment with human and social welfare.

Acknowledgments

This work was supported by the Research Council of Norway, TRUST-The Norwegian Centre for Trustworthy AI, project number 357425. We also gratefully acknowledge partial support from

NSF grant DMS-2413265, NSF grant DMS 2209975, NSF grant DMS-2515767, NSF grant 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI), the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639, and NSF grant MC2378 to the Institute for Artificial CyberThreat Intelligence and OperatioN (ACTION).

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[2] All Things Nordic. 2026. *Norway Draws a Line on AI in Schools.* Accessed September 2026. https://all-things-nordic.com/2026/06/22/norway-draws-a-line-on-ai-in-schools/

[3] Richard Antonello, Chandan Singh, Shailee Jain, Aliyah Hsu, Sihang Guo, Jianfeng Gao, Bin Yu, and Alexander Huth. 2024. Generative causal testing to bridge data-driven models and scientific theories in language neuroscience. *arXiv preprint arXiv:2410.00812* (2024).

[4] Kenneth A. Bamberger, Deirdre K. Mulligan, and Victor Zhenyi Wang. 2027. Contestability: epistemically-responsible knowledge production in AI governance. *Theoretical Inquiries in Law,* 28 (2027). (manuscript).

[5] Sumanta Basu, Karl Kumbier, James B Brown, and Bin Yu. 2018. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences* 115, 8 (2018), 1943–1948.

[6] Merle Behr, Yu Wang, Xiao Li, and Bin Yu. 2022. Provable Boolean interaction recovery from tree ensemble obtained via random forests. *Proceedings of the National Academy of Sciences* 119, 22 (2022), e2118636119.

[7] Landon Butler, Abhineet Agarwal, Justin Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. 2026. Proxyspex: Inference-efficient interpretability via sparse feature interactions in llms. *Advances in Neural Information Processing Systems* 38 (2026), 72306–72340.

[8] Dara-Abasi Ita, Forbes. 2026. *The Algorithm That Counted On No One Appealing.* Accessed September 2026. https://www.forbes.com/sites/daraabasiita/2026/06/09/the-algorithm-that-counted-on-no-one-appealing/

[9] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 31.

[10] Governor Gavin Newsom. 2026. *Governor Newsom signs first-in-the-nation AI safeguards to protect Californians, calls on the federal government to do its part.* Accessed September 2026. https://www.gov.ca.gov/2026/09/09/governor-newsom-signs-first-in-the-nation-ai-safeguards-to-protect-californians-calls-on-the-federal-government-to-do-its-part/

[11] Aliyah Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Odisho, Peter Carroll, and Bin Yu. 2025. Efficient automated circuit discovery in transformers using contextual decomposition. In *International Conference on Learning Representations*, Vol. 2025. 57880–57898.

[12] Aliyah R. Hsu, Yeshwanth Cherapanamjeri, Anobel Y. Odisho, Peter R. Carroll, and Bin Yu. 2024. Mechanistic Interpretation through Contextual Decomposition in Transformers. *arXiv preprint arXiv:2407.00886v1* (2024). arXiv:2407.00886v1 [cs.AI]

[13] Business Insider. 2026. *ChatGPT expands its foothold in the nation's classrooms.* Accessed September 2026. https://www.businessinsider.com/openai-chatgpt-for-teachers-us-schools-2026-8

[14] Justin Singh Kang, Landon Butler, Abhineet Agarwal, Yigit Efe Erginbas, Ramtin Pedarsani, Kannan Ramchandran, and Bin Yu. 2025. SPEX: Scaling feature interaction explanations for LLMs. *arXiv preprint arXiv:2502.13870* (2025).

[15] Karl Kumbier, Sumanta Basu, Erwin Frise, Susan E Celniker, James B Brown, and Bin Yu. 2018. Signed iterative random forests to identify enhancer-associated transcription factor binding. *arXiv preprint arXiv:1810.07287* (2018).

[16] Aaron J Li, Nicolas Sanchez, Hao Huang, Ruijiang Dong, Jaskaran Bains, Katrin Jaradeh, Zhen Xiang, Bo Li, Feng Liu, Aaron Kornblith, et al. 2026. Green Shielding: A User-Centric Approach Towards Trustworthy AI. *arXiv preprint arXiv:2604.24700* (2026).

[17] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4, 7 (2023).

[18] Deirdre K Mulligan, Daniel Kluttz, and Nitin Kohli. 2019. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. *Available at SSRN 3311894* (2019).

[19] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.

[20] NextGovFCW. 2026. *AI is already on the job, and Census is keeping count.* Accessed September 2026. https://www.nextgov.com/artificial-intelligence/2026/08/ai-already-job-and-census-keeping-count/415528/

[21] NHS England. 2026. *500,000 NHS staff to get new artificial intelligence tools to help free up more time for patients.*

Accessed September 2026. https://www.england.nhs.uk/2026/06/500000-nhs-staff-to-get-new-artificial-intelligence-tools-to-help-free-up-more-time-for-patients/

[22] Zachary T Rewolinski and Bin Yu. 2025. Pcs workflow for veridical data science in the age of ai. *arXiv preprint arXiv:2508.00835* (2025).

[23] Tiffany Tang. 2023. *Statistical Machine Learning for Reliable Hypothesis Generation in Biomedical Problems*. Ph. D. Dissertation. University of California, Berkeley.

[24] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature medicine* 26, 8 (2020), 1229–1234.

[25] United States District Court, S.D. New York. 2023. Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023).

[26] U.S. Census Bureau. 2026. *AI use at U.S. businesses*. Accessed September 2026. https://www.census.gov/library/stories/2026/05/ai-use-businesses.html

[27] Kristen Vaccaro, Karrie Karahalios, Deirdre K Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in algorithmic systems. In *Companion publication of the 2019 conference on computer supported cooperative work and social computing*. 523–527.

[28] Qianru Wang, Tiffany M Tang, Michelle Youlton, Chad S Weldy, Ana M Kenney, Omer Ronen, J Weston Hughes, Elizabeth T Chin, Shirley C Sutton, Abhineet Agarwal, et al. 2025. Epistasis regulates genetic control of cardiac hypertrophy. *Nature Cardiovascular Research* 4, 6 (2025), 740–760.

[29] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2369–2380.

[30] Bin Yu. 2013. Stability. *Bernoulli* (2013), 1484–1500.

[31] Bin Yu and Rebecca L Barter. 2024. *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press.

[32] Bin Yu and Karl Kumbier. 2018. Artificial intelligence and statistics. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 6–9.

[33] Bin Yu and Karl Kumbier. 2020. Veridical data science. *Proceedings of the National Academy of Sciences* 117, 8 (2020), 3920–3929.