PROCESSING, PRIMING, AND PROBING: HUMAN INTERVENTIONS FOR EXPLAINABILITY ALIGNMENT

Kenza Amara
Department of Computer Science
ETH Zurich, Switzerland
kenza.amara@ai.ethz.ch

ABSTRACT

As artificial intelligence (AI) systems play a central role in decision-making, the need for explainability becomes more critical. Effective explanations must balance two key objectives: faithfully representing the model's behavior while remaining reasonable and useful to humans. This dual requirement makes alignment a fundamental challenge in explainable AI (XAI). Research in human-centered XAI (HCXAI) has introduced guidelines and evaluation methods to enhance the accessibility and usability of explanations. These efforts have led to concrete strategies for incorporating human prior knowledge in the explainability pipeline. However, prioritizing human-centricity often comes at the cost of accurately reflecting the model's reasoning, behavior, and internal functioning. In this paper, we rigorously define *explainability alignment*, ensuring explanations remain both model- and human-centric without sacrificing one for the other. To maintain this balance, we propose targeted human interventions that enhance interpretability while preserving the core objective of XAI: making black-box models more transparent. To structure these interventions, we present *the Processing, Priming, and Probing (PPP) framework*, which categorizes different intervention strategies for achieving explainability alignment. They encompass (1) modifications to final explanations, (2) prior adjustments within a fixed XAI pipeline, and (3) novel approaches to designing and refining explanations with human supervision. Equipping researchers with such a framework will facilitate the development of more aligned explainability methods.

1 INTRODUCTION

As artificial intelligence (AI) systems become increasingly integrated into high-stakes decision-making processes and advances into the realm of personalization, they introduce new challenges (Kirk et al., 2024) that underscore the need for a deeper understanding of model behavior. Explainable AI (XAI) aims to provide insights into model behavior, but a key challenge persists: ensuring that explanations align with both the model's reasoning and human expectations.

Existing research in human-centered XAI (HCXAI) has explored various strategies for generating explanations that are more understandable and accessible to users. Prior work has focused on designing user-centered explanation frameworks (Liao & Varshney, 2021; Baber et al., 2024; Ehsan et al., 2023) and evaluating explainability methods from a human perspective (Kim et al., 2024; Colin et al., 2022; Nauta et al., 2023). Sanneman & Shah (2022) introduce the SAFE-AI framework, incorporating concepts from human factors literature—such as situation awareness, workload, and trust—into both the design and evaluation of XAI systems. Additionally, foundational research in interpretable machine learning (iML) has established criteria for assessing interpretability with a strong emphasis on human perception and understanding (Doshi-Velez & Kim, 2017). While these efforts have made valuable contributions, they primarily focus on defining human-centered requirements for XAI and quantifying human-centricity in XAI, with only a few suggestions for concrete strategies, lacking a unified framework of human interventions.

Parallel to HCXAI, research in informed machine learning has introduced taxonomies to integrate scientific knowledge into AI models to improve prediction accuracy (Von Rueden et al., 2021). This

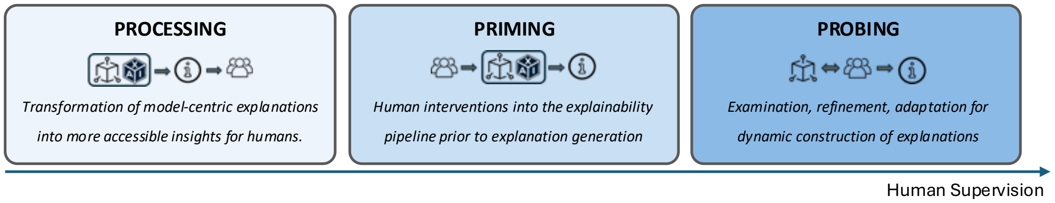
PROCESSING
Transformation of model-centric explanations into more accessible insights for humans.
PRIMING
Human interventions into the explainability pipeline prior to explanation generation
PROBING
Examination, refinement, adaptation for dynamic construction of explanations
Human Supervision

Figure 1: Summary of *the Processing, Priming, and Probing framework*. Human interventions to align model-centric explanations happen at different stages in the explainability pipeline, post-hoc, prior, or during the explanation generation process, with different levels of human supervision.

approach has been extended to XAI by Beckh et al. (2021) that propose a framework to incorporate prior knowledge into the ML pipeline or explainability methods to enhance interpretability. However, it predominantly focuses on enhancing explanations with prior knowledge to make them more accessible and contextualized (Beckh et al., 2021), rather than ensuring alignment between meeting human expectations and reflecting the model's behavior.

This raises a fundamental issue: explanations that appear intuitive, convincing, and useful may not necessarily reflect the model's actual reasoning process (Agarwal et al., 2024). Recent advances in large language models (LLMs) illustrate this challenge, as generative models can produce highly plausible rationales that align with human expectations but do not correspond to their internal decision-making processes (Agarwal et al., 2024).

To address the issue of plausible but inaccurate explanations, we must ensure that explanations remain faithful to the model while enhancing interpretability. This paper introduces *explainability alignment*, a concept that ensures explanations have both aspects from model- and human-centric explainability, capturing key properties from both perspectives. Unlike HCXAI, which prioritizes human interpretability, aligned explanations retain factual correctness by remaining grounded in the model's actual reasoning process. To implement explainability alignment in practice, we introduce *the Processing, Priming, and Probing (PPP) framework*. This framework classifies human interventions in XAI into three key types based on the required level of human supervision, as well as the timing and function of the intervention within the XAI pipeline, as illustrated in Figure 1. *Processing* refers to post-hoc modifications that transform generated explanations to make them more accessible for humans. *Priming* involves embedding prior knowledge into the explainability pipeline—whether at the model level or within explainability methods—to align explanations with domain expertise, scientific principles, and human-based rules. *Probing* extends beyond predefined pipelines, allowing for the co-creation of explanations through interactive mechanisms and new explainability designs. While prior research has explored priming explanations by incorporating prior knowledge into the XAI pipeline to improve explanation accessibility (Beckh et al., 2021), our framework goes further. It introduces intervention strategies for post-hoc refinement and new explanation formats, without the limitations of predefined XAI methods. Additionally, it classifies interventions based on when and how they occur during explanation generation. For example, while Beckh et al. (2021) group post-hoc concept discovery (Leemann et al., 2023), concept validation (Kim et al., 2018), and concept datasets for fine-tuning (Chen et al., 2020) under the broad category of informed machine learning, the PPP framework treats these as distinct types of interventions playing different roles in influencing explainability. Our contributions are:

- We characterize model- and human-centric explanations, identifying their distinct roles in XAI.
- We define explainability alignment as the goal of creating explanations that integrate both human-centric and model-centric aspects, ensuring the reasonableness of explanations while still faithfully representing the model's reasoning.
- We propose the PPP framework, which categorizes human interventions that refine model-centric explanations to enhance explainability alignment.

The PPP framework does not provide an exhaustive review of all human-centered alignment strategies but rather equips researchers with a structured classification of human interventions to facilitate alignment in explainability. Note that it excludes classic probing methods (Belinkov, 2022) and concept integration as done by concept-based XAI methods (Kim et al., 2018; Koh et al., 2020; Zarlenga et al., 2022). The rationale behind these choices is detailed in Appendix B.

2 Related Work

The AI Alignment Problem The development of artificial general intelligence (AGI) has immense potential, but also introduces significant risks, particularly the alignment problem. This challenge involves ensuring that AI systems pursue goals aligned with human values and interests rather than unintended or harmful objectives (Ngo et al., 2022; Hendrycks et al., 2020; Gabriel, 2020; Perez et al., 2022). The origins of the alignment problem trace back to the early days of AI research, with Norbert Wiener warning in 1960 that we must be certain the machine's purpose aligns with our own (Wiener, 1960). AI value alignment is essential before real-world deployment to ensure AI is aligned with human values (Russell, 2019), prioritizing values like capability, equity, and responsibility while avoiding harmful ones, such as seeking power to harm others (Ngo et al., 2022). Various definitions of alignment have emerged, focusing on human goals (Zhuang & Hadfield-Menell, 2020), preferences (Stray, 2020), or ethical principles (Irving & Askell, 2019).

Explainability Explainable artificial intelligence or interpretable machine learning have become a top priority in AI research. The goal of XAI is to enable users to "understand, appropriately trust, and effectively manage [...] artificially intelligent partners" (Gunning & Aha, 2019). Introduced by (Van Lent et al., 2004), XAI initially referred to a system's ability to explain AI-controlled entities' behavior in a U.S. Army training system. Since then, 36 terms related to XAI have been introduced (Vilone & Longo, 2021). Van Lent et al. (2004) defined XAI as presenting the user with "an easily understood chain of reasoning from the user's order, through the AI's knowledge and inference, to the resulting behavior". Explanations should be "easily understood", a point also emphasized by Biran & Cotton (2017) and Montavon et al. (2018), who state that systems are interpretable if their operations are understandable to humans. Nauta et al. (2023) frame explanations in XAI as representing the model's reasoning, functioning, and behavior in human-understandable terms, where reasoning is the process on how a model came to a particular decision, behavior only refers to how the model globally operates, and functioning refers to the (internal) workings and internal data structures of the machine learning models (Gilpin et al., 2018) [1].

3 Explainability Alignment

3.1 Definitions

While AI system alignment is essential for safety and ethical considerations, XAI Alignment is fundamental to the very existence of explainability. Despite the rapid development of XAI algorithms in recent years, these algorithms often fall short of how humans naturally produce and interpret explanations. As a result, many current XAI techniques are difficult to use and lack effectiveness. Misaligned explanations can lead to confusion, false confidence, or mistrust, ultimately undermining decision-making (Mueller et al., 2021; Ma et al., 2024). Providing meaningful and actionable explanations is a prerequisite for deploying explainable AI systems in real-world settings. Research has shown that when users develop accurate mental models of AI decision boundaries, they make more informed and effective AI-assisted decisions (Prasad et al., 2020). Achieving alignment between model explanations and human perspectives is therefore crucial for fostering understanding and trust in AI predictions [2]. *Explainability alignment* is the pursuit of explanations that are both:

- Consistent with the model's behavior – Explanations should accurately capture the model's decision-making process, ensuring the reproducibility of predictions and reflecting performance variations. They may be derived from model-aware explainability methods or external AI-based tools. Explanations that align with model behavior are referred to as *model-centric*.
- Consistent with human expectations – Explanations are reasonable, i.e., match ground truth, adhere to human rules, or are perceived as plausible [3] and useful, i.e., actionable and effective for decision-making. Explanations that align with human expectations are referred to as *human-centric*.

[1] Since these three aspects–reasoning, functioning and behavior– are closely interconnected in how models work, and their differentiation is specific to the explainability method, we use the term "behavior" to encompass all model actions in this paper.

[2] Research on human-aligned XAI often falls under the broader human-centered XAI literature, where scholars explore ways to make model-generated explanations more human-centric.

[3] Plausibility characterize explanations that meet human intuition and experience.

3.2 MODEL-CENTRIC EXPLAINABILITY

Model-centric explainability is poorly defined in the literature. We propose key aspects to characterize it [4]. *Model-centric explanations* focus on the internal mechanisms or decision-making processes of an AI model. They aim to provide insights into how the model arrives at its outputs by examining aspects such as feature importance, activation patterns, or decision rules. We identify four scenarios where explanations can be characterized as model-centric: (1) they have been generated by a model-aware explainability method, involving the model predictive power, sensitivity, gradient, or attention, (2) they are faithful to the initial model's behavior allowing reproducibility of predictions, (3) they align with some expected changes in the model's behavior after input perturbations, or (4) they are generated with the use of the model itself.

Model-Aware Explainability While model-aware explainability methods target the internal workings of a specific model, they may incorporate elements of both gradient, activation, and attention mechanisms. While they all aim to enhance transparency, they vary in terms of complexity and applicability. *Gradient-based methods* assess feature importance by analyzing how a model's output changes in response to its inputs. *Activation-based methods* analyze hidden layer activations to map feature importance to the input space. *Attention-based methods* analyze attention weights to reveal which input elements a model focuses on during prediction. Together, these methods form a comprehensive approach to understanding and interpreting the behavior of deep neural networks. We refer to subsection A.2 for more details about those methods.

Reproducibility of Predictions Faithful explanations—those that lead to the same prediction as the initial input—are model-centric. Faithfulness, a widely used model-based evaluation metric (Jacovi & Goldberg, 2020), assesses how well an explanation supports the original prediction by reintroducing it as input. Faithfulness metrics include keep-based (fidelity+) and removal-based (fidelity–) approaches (Yuan et al., 2020): fidelity+ retains explanatory elements, expecting the same prediction, while fidelity– removes them to measure prediction shifts. However, these metrics face limitations such as the out-of-distribution (OOD) problem, where removing entities can create inputs outside the training distribution, leading to unexpected model behavior. Ensuring model-centric explanations requires mitigating OOD effects, e.g., through retraining strategies (Hooker et al., 2018).

Perturbation-based Explainability Sensitivity to input perturbations (Ivanovs et al., 2021) represents an alternative method for verifying the model-centricity of explanations. Unlike faithfulness, which uses predefined explanatory entities as a basis for input modifications, this approach applies diverse perturbations to steer the model toward a desired behavior. For instance, in text-based tasks, small alterations of words can dramatically change the meaning of a sentence and therefore the output of a model (Liu et al., 2018). In image data, modifying specific pixel regions helps identify which features influence the model's behavior (Yang et al., 2021). The output variations are the objective and model alignment is reached by finding the right perturbations to meet that objective. The objective of perturbation-based XAI methods is ultimately to generate model-centric explanations.

Model as Tool for XAI In the strategy of using the model as an explainability agent, the model—typically a large language model—is tasked with generating its own rationale or explanation for its predictions and assesses the quality of its responses (Trivedi et al., 2025). Acting as an internal judge, it may justify its outputs, self-correct, or evaluate its reasoning (Gu et al., 2024). However, this approach is controversial, raising concerns about reliability. While fluent, their explanations may lack true understanding, as they rely on training data patterns rather than genuine reasoning (Agarwal et al., 2024; Ye & Durrett, 2022; Laban et al., 2023; Valmeekam et al., 2022).

3.3 HUMAN-CENTRIC EXPLAINABILITY

While significant research in HCXAI has identified key evaluation criteria for user-centric or human-centered explanations (Rong et al., 2023), we define human-centricity in XAI through the lens of alignment, focusing on pre-evaluation aspects. Specifically, while user-centricity depends on user perception, human-centricity stems from an explanation's alignment with human expectations—meaning that *human-centric explanations* should be both reasonable and useful (Colin et al., 2022). *Reasonableness* pertains to explanations that resonate with human cognitive processes. It

[4] As this is the first attempt to characterize model-centric XAI, we invite researchers to further refine and expand that list.

depends on the rationality of human expectations—whether based on truth (grounded in facts) or intuition (shaped by beliefs and experience). *Utility* or *usefulness* ensures that explanations are relevant to the user's needs and expertise level (Miller, 2019; Sokol & Flach, 2019). Detailed definitions of those two properties can be found in subsection A.1. In this section, we define three core criteria for human-centric explanations: (1) alignment with ground truth, where explanations match established, verifiable explanations, (2) adherence to human rules, where explanations follow domain-specific, linguistic, or logical principles, and (3) plausibility, where explanations reflect human intuition, beliefs, and desired outcomes.

Ground Truth Explanations Synthetic ground truth explanations provide valuable insights into what the model should be looking at for predictions. These explanations are constructed based on predefined key structures in the data, that we know are the reason for the predictions. These explanations allow researchers to assess how well models explain their predictions. Tools like GraphWorld (Tsitsulin et al., 2022), ERASER (DeYoung et al., 2020), and Virtual Synthetic Data (Man & Chahl, 2022) contribute significantly to advancing research on explainability by offering consistent, traceable ground truth explanations for various tasks. See subsection A.3 to learn more about ground truth explanations types and limitations.

Adherence to Rules Human-centric explanations should align with universally accepted rules, including scientific knowledge, linguistic rules, and logical reasoning. *Scientific knowledge* provides a robust foundation, drawing from mathematical proofs, physical laws, and biological observations. In various domains, experts offer ground truth explanations supported by scientific evidence, as seen in molecular datasets like MUTAG (Debnath et al., 1991), MoleculeNet (Wu et al., 2018), and Enzymes (Borgwardt et al., 2005). Adhering to *linguistic rules*, such as syntax—the structure of language to ensure clarity and logical flow—and grammar—rules guiding the proper use of tense, number, articles, and pronouns—, ensures clarity and logical flow, making explanations more accessible. Finally, *logical rules*, including deductive and inductive reasoning, structure relationships between concepts, leading to more consistent and predictable explanations. Human reasoning, based on formal systems like syllogisms, can be used to validate or clarify explanations.

Plausibility Plausible explanations are human-centric, deriving from human experience and intuition (Jacovi & Goldberg, 2020). They measure how convincing an explanation appears to humans and offer an intuitive way to assess whether an AI-generated output "makes sense". With intelligible modalities like images and text, these explanations leverage the low cognitive load such formats present. *Semantic intuition* relies on the meanings humans assign to words, phrases, and actions within context, making implausible explanations—such as a grammatically correct but semantically incorrect sentence—easy to detect. Beyond commonsense reasoning, it is shaped by experience, culture, and situational understanding. *Logical intuition* stems from prior expectations about how models should process familiar content, such as recognizing objects in images or detecting sentiment in text. These expectations enable quick validation of AI explanations with minimal cognitive effort. While semantic and logical intuition facilitate rapid assessment of an explanation's soundness, such human-centric explanations remain the least reliable and trustworthy (Jin et al., 2024).

4 The Processing, Priming, and Probing Framework

Explainability alignment aims to produce explanations that integrate both model-centric and human-centric perspectives, capturing the model's behavior while meeting human expectations. We specifically explore how human interventions can facilitate this balance. We propose the Processing, Priming, and Probing framework encompassing the diverse types of human interventions to align model-centric explanations. All interventions preserve the main purpose of explanations, namely representing model's behavior.

4.1 Overview

Given model-centric explanations defined in subsection 3.2, the Processing, Priming, and Probing framework, illustrated in Figure 2, examines external human interventions designed to enhance their alignment by incorporating more human-centric aspects from subsection 3.3. *Processing* encompasses all post-hoc interventions that transform generated explanations a posteriori to improve their comprehensibility. This stage focuses on refining existing explanations without modifying the un-

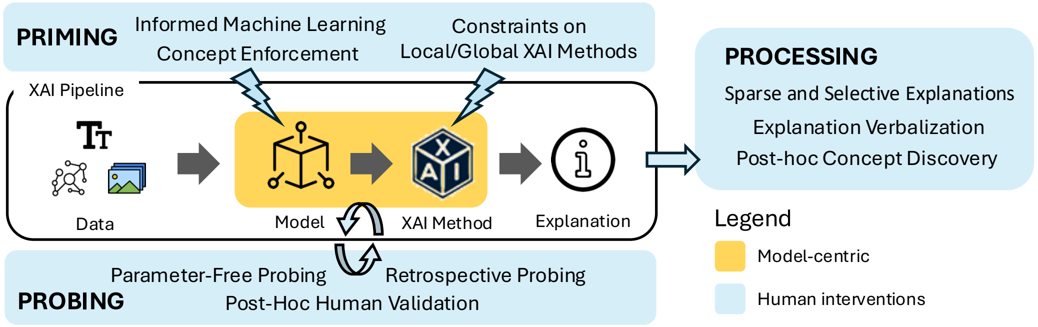

Figure 2: The Processing, Priming, and Probing Framework. How and where in the XAI pipeline humans intervene.

derlying model or the explainability pipeline. In contrast, *Priming* refers to interventions applied at an earlier stage of the explanation generation process. These interventions modify the model's objective function, optimize the eplainability method, or introduce constraints such as regularization to incorporate prior knowledge—such as scientific principles or logical rules—into the model's learning process [5]. *Probing* interventions, on the other hand, require substantially greater human supervision. These methods involve defining data-centric controlled tasks to assess the model's understanding, collecting human feedback to refine explanations, and validating alignment between explanations and expected reasoning patterns. Probing can also introduce alternative forms of explanation by systematically testing the model's sensitivity.

4.2 Processing Interventions

Processing represents the simplest strategy for transforming model-centric explanations into more understandable and actionable insights for humans. By shaping sparse and selective explanations, verbalizing feature-based attributions, and leveraging conceptual abstraction, post-hoc human interventions enhance the accessibility of existing model-centric explanations.

Towards Sparse and Selective Explanations A key challenge in explanation design is ensuring simplicity and intelligibility. Limiting explanation size improves clarity. Explanations sparsity strategies address this issue by filtering only the most relevant features (Amara et al., 2022). Another factor is whether explanations should be weighted or binary, determining if importance ranking is needed or whether all important entities should be treated equally. Additionally, Lai et al. (2023) show that effective explanation communication depends on selecting information based on the explainer's goal and beliefs about the recipient. Following this observation, they propose a selective explanation framework to adjust AI explanations based on user preferences, focusing on relevance, abnormality, and changeability to enhance perceived understanding.

Explanation Verbalization Making explanations accessible also involves translating technical outputs into user-friendly narratives. Natural language rationalization improves user engagement by converting complex reasoning into intuitive text (Ehsan et al., 2018). Expanding beyond rationalization, research on verbalization and visualization demonstrates the benefits of combining multiple modalities to enhance explainability (Sevastjanova et al., 2018). Feldhus et al. (2022) demonstrate how saliency map verbalization reduces cognitive load, making explanations easier to comprehend compared to conventional heatmap visualizations. Rong et al. (2024) extract visual feature maps from the classifier with an attention module and generate descriptive sentences. Another approach leverages LIME in combination with Inductive Logic Programming to generate verbal explanations for image classification (Rabold et al., 2020).

Post-hoc Concept Discovery Beyond transforming individual explanations, adapting explanations to user cognitive models is another crucial aspect of accessibility. Traditional saliency- or attribution-based explanation techniques, while widely used, highlight the importance of specific input regions but do not clarify what these regions represent in terms of human-understandable concepts (Kim et al., 2018). In recent years, post-hoc concept discovery has emerged as a powerful technique for aligning explanations with human reasoning (Leemann et al., 2023). A particularly

[5] Unlike Beckh et al. (2021), we do not consider human feedback as a kind of prior knowledge.

promising approach in this domain is concept relevance propagation (CRP), which bridges local and global perspectives in explainability (Achtibat et al., 2023). CRP integrates attribution-based local explanations with global concept-level representations using relevance maximization to generate explanations that are both human-interpretable and faithful to the model's learned representations.

4.3 PRIMING INTERVENTIONS

A second approach to enhancing explanation alignment is priming [6]. Priming explanations involve introducing human interventions into the explainability pipeline before the explanations are generated, reinforcing human-centricity in the final outputs. This can occur either by priming the AI model, shaping its training process to yield more reasonable explanations, or by priming the XAI method, constraining the explainability algorithm with human-based rules. For a detailed review of methods that integrate prior knowledge into training data, modify ML architectures, or apply regularization techniques in learning algorithms, we refer to Beckh et al. (2021).

4.3.1 PRIMING THE MODEL

One effective way to align model explanations with human understanding is by integrating prior human knowledge into the model's learning objectives. This can be achieved by modifying the loss function, adding regularization terms, or enforcing scientific constraints.

Informed Machine Learning In the domain of physics-informed machine learning, researchers have explored various strategies for integrating physical principles into ML models (Von Rueden et al., 2021). These include physics-guided loss functions, physics-aware initialization, physics-constrained architecture design, and hybrid modeling. One of the most common techniques for maintaining consistency with physical laws is incorporating domain-specific constraints into the loss function (Karpatne et al., 2017). For instance, in applications such as compound activity prediction, loss functions can be adapted to enforce prior about the molecular structure (Amara et al., 2023). By embedding scientific principles as additional loss terms, the model's behavior remains aligned with real-world phenomena throughout its training and model-aware explainability methods, consequently leading to more aligned explanations.

Concept Enforcement An alternative model-level intervention is concept enforcement, where ML models are explicitly trained to align with human-interpretable concepts. Chen et al. (2020) propose replacing standard batch normalization layers in neural networks with concept whitening layers, which decorrelate input features before aligning them with predefined human concepts. For example, in image classification tasks, a convolutional neural network (CNN) trained with concept whitening layers can be fine-tuned using an external dataset labeled with interpretable concepts such as "airplane" or "person". This alignment process not only helps debug model training—by detecting misalignment between similar concepts—but also enhances the interpretability of decision processes, as the model's predictions can be decomposed into recognizable conceptual components.

4.3.2 PRIMING THE EXPLAINABILITY METHOD

Instead of modifying the model itself, an alternative approach is to constrain the explainability method directly, ensuring explanation alignment. This involves modifying the optimization processes of explainability techniques, such as activation maximization, gradient-based methods, or coalition-based approaches, to balance model-awareness and human interpretability.

Interventions on Local Explainability Methods Several studies have proposed explainability techniques that incorporate human rationales (Ehsan et al., 2019). These methods identify key linguistic features, enforce logical constraints, or adhere to syntactic and grammatical rules to produce more intuitive explanations. For example, Ehsan et al. (2019) trained an explanation generation model using human rationale data to assist non-expert users in interpreting model behavior. Similarly, Feng & Boyd-Graber (2022) developed a model that selects tailored explanations based on different user preferences. Amara et al. (2024b) introduced a coalition-based explainability method that integrates syntactic rules into the widely used SHAP framework to generate human-compliant explanations

[6] In psychology, priming refers to how exposure to one stimulus can unconsciously influence responses to a subsequent stimulus (Bargh & Chartrand, 2000).

in next-token prediction tasks for LLMs. By ensuring that explanations align with syntactic dependencies in the input sentence, this method enhances the interpretability of LLM decoder token generation.

Interventions on Global Explainability Methods Research has also explored interventions on global explainability methods such as activation maximization (AM), which seeks to discover the optimal input patterns that maximize a model's activation for a particular class. Enhancements to AM have introduced additional algebraic constraints to improve interpretability. For instance, Mahendran & Vedaldi (2016) constrained the total variation of explanations by anchoring them to prior image distributions, producing smoother visual outputs. Similarly, Yosinski et al. (2015) penalized high-frequency artifacts in activation-based visualizations by applying Gaussian blur kernels at each optimization step. Beyond AM, knowledge graphs (KGs) have proven valuable for concept-based explainability methods (Lecue, 2020; Longo et al., 2024). KGs have been used to define concepts for concept bottleneck models, enabling models to reason about concepts without explicit labeled supervision (Oikarinen et al., 2023; Yuksekgonul et al., 2022). By leveraging structured knowledge representations, these methods enhance both the interpretability and truthfulness of explanations.

4.4 Probing Interventions

The third approach is probing which comprises interventions for XAI alignment that require greater human supervision. Probing explanations refers to the systematic process of examining, refining, and adapting explanations to better match human expectations. Unlike conventional XAI methods, where explanations are derived from external explainability techniques, this approach dynamically constructs explanations as part of the probing process. It relies on targeted perturbations, human validation, and feedback-driven refinement to generate aligned explanations, both reasonable and reflective of the model's behavior [7].

4.4.1 Parameter-free Probing

Semantic perturbations modify inputs in a controlled way to align explanations with human expectations of model behavior. These parameter-free probing methods [8] (Zhao et al., 2024a) rely on datasets designed to evaluate specific properties, such as grammar (Marvin, 2018), and assess how well a model encodes these properties based on its performance. Well-designed tasks should yield predictable performance shifts, as seen in CheckList (Ribeiro et al., 2020), which guides users in testing linguistic capabilities. Counterfactually-augmented data (Kaushik et al., 2019) refine inputs with human-in-the-loop edits, e.g., flipping a review's sentiment with minimal changes. CausalGym (Arora et al., 2024) benchmarks interpretability by identifying causal linguistic features. In multimodal settings, Amara et al. (2024a) explore perturbations in visual question answering. The dynamic nature of perturbations helps users build intuitions, form hypotheses, and test them instantly. These human expectations enable to shape controlled, meaningful interventions to reach explainability alignment.

4.4.2 Post-Hoc Human Validation

A straightforward method for probing explanations to improve alignment is to measure human satisfaction by comparing the raw outputs of XAI algorithms against human rationales.

Alignment Metric Validation Abstraction alignment, introduced by Boggust et al. (2024), provides a methodology for assessing the agreement between a model's learned abstractions and human expectations, such as linguistic hierarchies or medical disease ontologies. The authors propose key metrics such as the "human-aligned" and the "sufficient subset" metrics, which evaluate respectively the frequency and extent to which the model's rationale aligns with human reasoning. Zhao et al. (2024b) propose using mutual information (MI) as an alignment measure. They treat a well-trained explanation generation model as a backbone, fine-tuning it further using reinforcement learning with

[7] The probing interventions in the PPP framework also incorporate feedback on the model's explanations to regularize the model's behavior towards the desired outcome, which is considered a form of prior knowledge in Beckh et al. (2021).

[8] Another category of data-centric probing techniques.

MI-based guidance. The MI estimator rewards generated explanations that are more aligned with predicted ratings or predefined features of recommended items.

Human Concepts in Model Explanations Although alignment is hard to quantify, concept-based explanations are often more intuitive for humans (Das et al., 2023) and generally more interpretable in classification tasks (Kim et al., 2018; Koh et al., 2020; Zarlenga et al., 2022; Ghorbani et al., 2019). Some approaches validate explanations against knowledge bases (Doran et al., 2017) or use external datasets to evaluate how well a model's latent representations align with predefined concepts [9] (Kim et al., 2018). Probing explanations and model representations for human concepts helps assess and enhance explainability alignment.

4.4.3 Retrospective Probing

When multiple plausible explanations exist without a clear alignment criterion—such as the absence of an alignment metric or justification for prioritizing one explanation over another—user studies are essential for evaluation (Doshi-Velez & Kim, 2018). Unlike post-hoc human validation, user studies capture subjective preferences, assessing factors like clarity, trust, decision-making guidance, and actionability (Singh et al., 2024). Various XAI studies focus on human performance in interpretability tasks (Narayanan et al., 2018) or compare model interpretability through A/B testing (Lakkaraju et al., 2016; Kim et al., 2014). Collected feedback refines explainability, serving as a supervision signal for model training (Carton et al., 2021; Stumpf et al., 2009; Ghai et al., 2021). For example, explainable active learning (XAL) allows annotators to critique explanations, improving model outputs (Ghai et al., 2021), while other studies use feedback to revise training data or modify learning algorithms (Beckh et al., 2021). Thus, user studies are key to aligning model-centric explanations with human preferences.

5 Discussion

We define explainability alignment as integrating both model- and human-centric aspects of explainability. To structure this, we introduce the PPP framework, categorizing human interventions based on supervision level, role, and timing in the XAI pipeline. This framework helps systematically approach explainability alignment.

Loss of Alignment While the PPP framework aligns model-centric explanations, excessive human-centricity can undermine XAI's goal of faithfully representing a model's reasoning. As noted in Agarwal et al. (2024), improving plausibility may reduce faithfulness—a phenomenon we call "Too Plausible to Be True". This issue is common in LLM-generated self-explanations (Kunz & Kuhlmann, 2024). Bommasani et al. (2021) and Sevastjanova & El-Assady (2022) warn us against misleading rationalizations. In addition, human supervision has limitations: Tan (2022) question whether human explanations should serve as additional supervision or as ground truth, as they often contain errors. Models tend to mimic human misconceptions (Lin et al., 2021), which can further distort generated explanations. Excluding human input from evaluation may sometimes be preferable to focus on the model's true reasoning (Petsiuk, 2018).

Quantifying XAI Alignment A persistent challenge is how to quantify explainability alignment. This issue extends beyond the AI alignment community to the HCXAI domain, where it largely revolves around human-centered evaluation. Recent surveys (Rong et al., 2023) highlight the ongoing need for more transparent and comparable human-based evaluations in XAI, since assessing the quality of human-centered activities remains difficult due to the complexity of the underlying concepts (Sankowski & Krause, 2023). Efforts to define metrics include frequency-based LLM prompting (Norhashim & Hahn, 2024) and Markov Decision Process modeling (Barez & Torr, 2023). But reducing human values to reward functions is inherently reductive (Ji et al., 2023) and defining moral values remains difficult (Awad et al., 2018). Beyond capturing human values, quantifying XAI alignment must also account for more nuanced, human-specific aspects of explainability, such as aligning with intuition, which remain difficult to measure universally.

[9] While Chen et al. (2020) fine-tune the model using external datasets for alignment, Kim et al. (2018) assess it post-training.

REFERENCES

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

Kenza Amara, Rex Ying, Zitao Zhang, Zhihao Han, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. Graphframex: Towards systematic evaluation of explainability methods for graph neural networks. *arXiv preprint arXiv:2206.09677*, 2022.

Kenza Amara, Raquel Rodríguez-Pérez, and José Jiménez-Luna. Explaining compound activity predictions with a substructure-aware loss for graph neural networks. *Journal of cheminformatics*, 15(1):67, 2023.

Kenza Amara, Lukas Klein, Carsten Lüth, Paul Jäger, Hendrik Strobelt, and Mennatallah El-Assady. Why context matters in vqa and reasoning: Semantic interventions for vlm input modalities. *arXiv preprint arXiv:2410.01690*, 2024a.

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. Syntaxshap: Syntax-aware explainability method for text generation. *arXiv preprint arXiv:2402.09259*, 2024b.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Gradient-based attribution methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 169–191, 2019.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv preprint arXiv:2402.12560*, 2024.

Pepa Atanasova. A diagnostic study of explainability techniques for text classification. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pp. 155–187. Springer, 2024.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

C Baber, P Kandola, I Apperly, and E McCormick. Human-centred explanations for artificial intelligence systems. *Ergonomics*, pp. 1–15, 2024.

Fazl Barez and Philip Torr. Measuring value alignment, 2023. URL `https://arxiv.org/abs/2312.15241`.

John A Bargh and Tanya L Chartrand. The mind in the middle. *Handbook of research methods in social and personality psychology*, 2:253–285, 2000.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

Katharina Beckh, Sebastian Müller, Matthias Jakobs, Vanessa Toborek, Hanxiao Tan, Raphael Fischer, Pascal Welke, Sebastian Houben, and Laura von Rueden. Explainable machine learning with prior knowledge: an overview. *arXiv preprint arXiv:2105.10172*, 2021.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3889–3900, 2022.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pp. 8–13, 2017.

Angie Boggust, Hyemin Bang, Hendrik Strobelt, and Arvind Satyanarayan. Abstraction alignment: Comparing model and human conceptual relationships. *arXiv preprint arXiv:2407.12543*, 2024.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1): i47–i56, 2005.

Samuel Carton, Surya Kanoria, and Chenhao Tan. What to learn, and how: Toward effective learning from rationales. *arXiv preprint arXiv:2112.00071*, 2021.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in neural information processing systems*, 35:2832–2845, 2022.

Devleena Das, Sonia Chernova, and Been Kim. State2explanation: Concept-based explanations to benefit agent learning and user understanding. *Advances in Neural Information Processing Systems*, 36:67156–67182, 2023.

Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Derek Doran, Sarah Schulz, and Tarek R. Besold. What does explainable ai really mean? a new conceptualization of perspectives, 2017. URL https://arxiv.org/abs/1710.00794.

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

Finale Doshi-Velez and Been Kim. Considerations for evaluation and generalization in interpretable machine learning. *Explainable and interpretable models in computer vision and machine learning*, pp. 3–17, 2018.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 81–87, 2018.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th international conference on intelligent user interfaces*, pp. 263–274, 2019.

Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai. *Proceedings of the ACM on human-computer interaction*, 7(CSCW1):1–32, 2023.

Nils Feldhus, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Möller. Constructing natural language explanations via saliency map verbalization. *arXiv preprint arXiv:2210.07222*, 2022.

Shi Feng and Jordan Boyd-Graber. Learning to explain selectively: A case study on question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–28, 2021.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. IEEE, 2018.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. Evaluating feature importance estimates. *arXiv preprint arXiv:1806.10758*, 2, 2018.

Geoffrey Irving and Amanda Askell. Ai safety needs social scientists. *Distill*, 4(2):e14, 2019.

Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Why is plausibility surprisingly problematic as an xai criterion?, 2024. URL https://arxiv.org/abs/2303.17707.

Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331, 2017.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Been Kim, Cynthia Rudin, and Julie A Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems*, 27, 2014.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Jenia Kim, Henry Maathuis, and Danielle Sent. Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence*, 7:1456486, 2024.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pp. 1–10, 2024.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Jenny Kunz and Marco Kuhlmann. Properties and challenges of llm-generated explanations. *arXiv preprint arXiv:2402.10532*, 2024.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander R Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. Llms as factual reasoners: Insights from existing benchmarks and beyond. *arXiv preprint arXiv:2305.14540*, 2023.

Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. Selective explanations: Leveraging human input to align explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–35, 2023.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1675–1684, 2016.

Freddy Lecue. On the role of knowledge graphs in explainable ai. *Semantic Web*, 11(1):41–51, 2020.

Tobias Leemann, Michael Kirchhof, Yao Rong, Enkelejda Kasneci, and Gjergji Kasneci. When are post-hoc conceptual explanations identifiable? In *Uncertainty in Artificial Intelligence*, pp. 1207–1218. PMLR, 2023.

Q Vera Liao and Kush R Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Shusen Liu, Zhimin Li, Tao Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. Nlize: A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. *IEEE transactions on visualization and computer graphics*, 25(1):651–660, 2018.

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. Rethinking attention-model explainability through faithfulness violation test. In *International Conference on Machine Learning*, pp. 13807–13824. PMLR, 2022.

Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

Jiaqi Ma, Vivian Lai, Yiming Zhang, Chacha Chen, Paul Hamilton, Davor Ljubenkov, Himabindu Lakkaraju, and Chenhao Tan. Openhexai: An open-source framework for human-centered evaluation of explainable machine learning. *arXiv preprint arXiv:2403.05565*, 2024.

Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120:233–255, 2016.

Keith Man and Javaan Chahl. A review of synthetic image data and its use in computer vision. *Journal of Imaging*, 8(11):310, 2022.

Rebecca Marvin. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*, 2018.

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

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

Shane T Mueller, Elizabeth S Veinott, Robert R Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J Clancey. Principles of explanation in human-ai systems. *arXiv preprint arXiv:2102.04972*, 2021.

T Nathan Mundhenk, Barry Y Chen, and Gerald Friedland. Efficient saliency maps for explainable ai. *arXiv preprint arXiv:1911.11293*, 2019.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Ernst Niebur. Saliency map. *Scholarpedia*, 2(8):2675, 2007.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

Hakim Norhashim and Jungpil Hahn. Measuring human-ai value alignment in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 1063–1073, 2024.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

V Petsiuk. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. To what extent do human explanations of model behavior align with actual model behavior? *arXiv preprint arXiv:2012.13354*, 2020.

Johannes Rabold, Hannah Deininger, Michael Siebers, and Ute Schmid. Enriching visual with verbal explanations for relational concepts–combining lime with aleph. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*, pp. 180–192. Springer, 2020.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

Yao Rong, David Scheerer, and Enkelejda Kasneci. Faithful attention explainer: Verbalizing decisions based on discriminative features. *arXiv preprint arXiv:2405.13032*, 2024.

Stuart Russell. *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

Olga Sankowski and Dieter Krause. The human-centredness metric: early assessment of the quality of human-centred design activities. *Applied Sciences*, 13(21):12090, 2023.

Lindsay Sanneman and Julie A Shah. The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems. *International Journal of Human–Computer Interaction*, 38(18-20):1772–1788, 2022.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Rita Sevastjanova and Mennatallah El-Assady. Beware the rationalization trap! when language model explainability diverges from our mental models of language. *arXiv preprint arXiv:2207.06897*, 2022.

Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel A Keim, and Mennatallah El-Assady. Going beyond visualization: Verbalization as complementary medium to explain machine learning models. 2018.

Ronal Singh, Tim Miller, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. An actionability assessment tool for explainable ai. *arXiv preprint arXiv:2407.09516*, 2024.

Kacper Sokol and Peter Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. December 2019.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Ramya Srinivasan and Ajay Chander. Explanation perspectives from the cognitive sciences—a survey. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 4812–4818, 2021.

Jonathan Stray. Aligning ai optimization to community well-being. *International Journal of Community Well-Being*, 3(4):443–463, 2020.

Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International journal of human-computer studies*, 67(8):639–662, 2009.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Chenhao Tan. On the diversity and limits of human explanations. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2173–2188, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.158. URL `https://aclanthology.org/2022.naacl-main.158/`.

Prapti Trivedi, Aditya Gulati, Oliver Molenschot, Meghana Arakkal Rajeev, Rajkumar Ramamurthy, Keith Stevens, Tanveesh Singh Chaudhery, Jahnavi Jambholkar, James Zou, and Nazneen Rajani. Self-rationalization improves LLM as a fine-grained judge, 2025. URL https://openreview.net/forum?id=RZZPnAaw6Z.

Anton Tsitsulin, Benedek Rozemberczki, John Palowitch, and Bryan Perozzi. Synthetic graph generation to benchmark graph learning. *arXiv preprint arXiv:2204.01376*, 2022.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

Michael Van Lent, William Fisher, and Michael Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pp. 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

Laura Von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, et al. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633, 2021.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

Norbert Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410): 1355–1358, 1960.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Qing Yang, Xia Zhu, Jong-Kae Fwu, Yun Ye, Ganmei You, and Yuan Zhu. Mfpp: Morphological fragmental perturbation pyramid for black-box model explanations. In *2020 25th International conference on pattern recognition (ICPR)*, pp. 1376–1383. IEEE, 2021.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392, 2022.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. *Adv. Neural Inf. Process. Syst.*, 32:9240–9251, December 2019.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. December 2020.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, et al. Concept embedding models. In *NeurIPS 2022-36th Conference on Neural Information Processing Systems*, 2022.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024a.

Yurou Zhao, Yiding Sun, Ruidong Han, Fei Jiang, Lu Guan, Xiang Li, Wei Lin, Weizhi Ma, and Jiaxin Mao. Aligning explanations for recommendation with rating and feature via maximizing mutual information. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3374–3383, 2024b.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.