5(0) Shades of Wrong: Disentangling the Wrongness of AI Explanations

LUCA DECK*, University of Bayreuth, Germany and Fraunhofer FIT, Germany
ANTON HUMMEL*, XITASO GmbH, Germany and University of Bayreuth, Germany
PAULA ZIETHMANN, Institute for Employment Research (IAB), Germany
KATELYN MORRISON, Carnegie Mellon University, USA
PHILIPP SPITZER, Karlsruhe Institute of Technology (KIT), Germany
NIKLAS KÜHL, University of Bayreuth, Germany and Fraunhofer FIT, Germany

The prevalence of imperfect AI explanations, whether due to dark patterns or unintentional explainability pitfalls, poses significant risks to human-AI collaboration. However, existing evaluation metrics and criteria for determining whether an explanation is wrong lack convergence on what it means for explanations to be "wrong". This paper proposes a typology conceptualizing the wrongness of explanations along three dimensions: *faithfulness* (reflecting the model's actual reasoning), *validity* (describing relationships with a ground truth), and *coherence* (maintaining an internal logic). We derive five explanation archetypes ("5 shades of wrong"): *Perfect Liar*, *Illuminator*, *Storyteller*, *Confused Illuminator*, and *Hallucinator*—illustrated through an image classification use case with heatmap explanations. We analyze potential risks and benefits of these archetypes and propose mitigation strategies for developers and users.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**.

Additional Key Words and Phrases: Explainable AI, AI Explanations, Dark Patterns, Explainability Pitfalls, XAI Literacy

ACM Reference Format:
Luca Deck, Anton Hummel, Paula Ziethmann, Katelyn Morrison, Philipp Spitzer, and Niklas Kühl. 2026. 5(0) Shades of Wrong: Disentangling the Wrongness of AI Explanations. In *Proceedings of MIRAGE Workshop at ACM Conference on Itelligent User Interfaces (IUI'26).* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Despite the increasing importance of explainability and transparency of artificial intelligence (AI), critical voices have been raised regarding the false promises, limitations, and misuses of Explainable Artificial Intelligence (XAI) (e.g., 4, 8, 13). Explanations that aim to illuminate or justify the inner workings of AI models have been linked to skewed depictions of the underlying models [2, 49], over-reliance or under-reliance on AI predictions [46], unwarranted confidence and trust [8, 27], or increased perpetuation of biases and stereotypes [41, 57]. The surfacing risks have been broadly categorized into deceptive patterns (DPs), where the limitations of XAI are intentionally exploited to

*Both authors contributed equally to this research.

Authors' Contact Information: Luca Deck, luca.deck@uni-bayreuth.de, University of Bayreuth, Bayreuth, Germany and Fraunhofer FIT, Bayreuth, Germany; Anton Hummel, anton.hummel@xitaso.com, XITASO GmbH, Augsburg, Germany and University of Bayreuth, Bayreuth, Germany; Paula Ziethmann, paula.ziethmann@iab.de, Institute for Employment Research (IAB), Nuremberg, Germany; Katelyn Morrison, kcmorris@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, USA; Philipp Spitzer, philipp.spitzer@kit.edu, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany; Niklas Kühl, niklas.kuehl@uni-bayreuth.de, University of Bayreuth, Bayreuth, Germany and Fraunhofer FIT, Bayreuth, Germany.

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.
Manuscript submitted to ACM

mislead or manipulate, and explainability pitfalls (EPs), where the limitations of XAI manifest accidentally due to a lack of awareness or transparency [13][1]. However, limitations are often inherent to the imperfect underlying models and imperfect XAI approaches that cannot be fully solved at a technical level [46]. This prompts the AI community to understand and communicate the limitations of XAI more precisely. As a response, a range of evaluation methods and criteria has been proposed to quantify the quality of explanations (e.g., 37, 45). Still, there is an ambiguity in the XAI research community about what it means for explanations to be right or wrong. For example, in the evaluation of a user study on imperfect XAI, Spitzer et al. [46] note: "[S]ome explanations that we classify as incorrect can contain correct evidence, making it difficult to have only a binary categorization for the correctness of explanations." Thus, in this paper, we expand upon a criterion-focused view by systematically analyzing dimensions of wrongness and propose *5 shades of wrongness* for a more nuanced understanding of how explanations can be wrong. The 5 shades of wrongness represent idealized archetypes that offer room for even more shades and combinations when considered as a spectrum of wrongness dimensions in practical applications. Crucially, this work does not aim to introduce merely another set of quality metrics; rather, it seeks to pragmatically navigate the realistic concession that explanations are rarely perfect by focusing on the sensitization to and prioritization of problematic types of wrongness. By delineating these archetypes, we support developers and auditors in prioritizing the mitigation of the most harmful risks of wrong explanations, while simultaneously sensitizing affected parties to the specific vulnerabilities arising from certain forms of wrongness.

After summarizing existing work on limitations and evaluations of XAI in Section 2, we introduce the three dimensions of wrongness: *faithfulness*, *validity*, and *coherence* in Section 3. Afterwards, we illustrate the different configurations of wrongness by introducing six explanation archetypes using a running example from credit applications in Section 4. With this typology, we shape a more nuanced understanding and vocabulary of wrongness in XAI and argue that certain configurations of wrongness carry higher risks than others (see Section 5). Looking ahead, we discuss strategies to deal with wrongness in Section 6 and plan to embed this typology within the larger endeavor of *XAI literacy*, which encompasses academic and training efforts aimed at increasing knowledge and awareness of XAI limitations among developers, users, and affected parties. Section 7 discusses the limitations of our work and concludes.

2 Related Work

Prior work has proposed holistic accounts of quality criteria [37] and desiderata stakeholders connect to XAI [28] as well as failures that can hinder desiderata satisfaction during the explanation process [5]. Starting from these overviews, the purpose of our work is to precisely define the ways in which explanations can be *wrong*. We therefore strictly focus on a certain type of system-specific limitations in the form of discrepancies between the explanation content and a "right" baseline, explicitly excluding user-specific expectations and interpretations. This augments Bove et al. [5]'s typology of system-specific failures which only superficially considers the triadic relationship between ground truth, model, and explanation. The following section will briefly summarize work on the multiple levels of discrepancies and the evaluation methods (EMs) that have been proposed to address those. Afterwards, we explain why—amidst the many positive attempts of evaluation—it is worthwhile to examine explanations from a wrongness-focused perspective.

2.1 Discrepancies of Explanations

Following Ehsan and Riedl [13], we categorize discrepancies into intentional *DPs* and unintentional *EPs*. The overview of related work in Table 1 illustrates that these discrepancies manifest at different levels, depending on which frame

[1]The term "deceptive patterns" is synonymous with "dark patterns", but selected here because it is more precise and avoids Black/White metaphors.

of reference the explanation is compared to. The majority of papers concern discrepancies *between the content of the explanation and the model behavior* it is supposed to represent. Intentional discrepancies in adversarial settings range from straight-up lying to addressees [24, 29] to deceptively forging intended explanations, e.g., by manipulating the underlying model [44] or exploiting correlations of features [2]. But even without intentional deception, explanations can misrepresent the underlying model [33] or be too complex for human addressees to comprehend [18, 54]. Moreover, studies have warned to look out for explanations that might be correct in and of themselves but irrelevant to the addressee, e.g., because they are disconnected from the underlying model [7, 14].

A second discrepancy can occur *between the explanation and the ground truth*. Straightforwardly, this can happen when the explanation faithfully represents an imperfect model [49]. In the more nuanced case, however, this discrepancy stems from imperfections within the explanation methods themselves. Post-hoc explainers often lack guarantees of correctness and can systematically misrepresent the ground truth, regardless of the underlying model's actual quality [4, 16, 46]. When explanations diverge from the ground truth, this can obscure actual model performance, lead to inappropriate levels of reliance, and ultimately compromise decision-making—especially for laypeople [7, 35]. Evaluating and guaranteeing the true validity of XAI methods against real-world observations remains a significant methodological challenge [31, 33].

Finally, discrepancies can occur *within or between explanations*. Internal discrepancies for a single explanation may be due to logical contradictions of rule-based [15] or large language model-based explanations [25]. "Unstable" explainers [5] yield differing explanations when applying the same explainer to identical or similar inputs. This can occur for XAI methods like LIME [40] which employ stochastic perturbations that can lead to unstable explanations [51, 56]. In cases where multiple explainers are used, recent literature has studied two phenomenon: the "disagreement problem" [26] for differing explanations generated from the same model; and the "Rashomon Effect" [36] for explanations of differing models with equal performance. Our categorization of discrepancies underscores that "wrongness" is not a monolithic concept, but a multifaceted risk requiring targeted mitigation strategies.

2.2 Evaluation Methods and Criteria

In light of the discrepancies outlined above, XAI research has been accompanied by a proliferation of EM and criteria for explanations. Several comprehensive surveys have documented this landscape (e.g., 37). Prior work on EMs can be organized along multiple dimensions, primarily distinguishing between different levels of abstraction: evaluation properties at the conceptual level, the manner in which explanations are assessed, the specific EMs themselves, and EM software tools.

The literature offers different perspectives on categorizing EMs. One fundamental question concerns *how* the XAI method is evaluated. Several taxonomies (e.g., 12, 50) distinguish between EMs that incorporate humans (qualitative) and those that do not (quantitative). The taxonomy by Doshi-Velez and Kim [12] distinguishes between application-grounded, human-grounded, and functionally-grounded evaluation. Application-grounded and human-grounded approaches incorporate humans in the evaluation process (e.g., *explanation satisfaction scale;* 19), whereas functionally-grounded evaluation relies on "formal definitions of interpretability as a proxy for explanation quality" [12, p.5]. Compared to qualitative evaluations, these quantitative approaches are considerably more time- and cost-efficient (e.g., *infidelity;* 55). However, while this taxonomy provides guidance on *how* to evaluate, it remains agnostic to characterizing the specific ways explanations can be wrong or misleading.

A second perspective addresses *what* the EM evaluates. For instance, [37] propose the Co-12 Properties, viewing interpretability as an inherently multi-faceted concept. They aim to standardize and improve XAI evaluation practices

Table 1. Overview of related work demonstrating or discussing different discrepancies of explanations, grouped according to types of discrepancies and whether they require intent (DPs) or can occur accidentally (EPs). Entries placed in the middle can be subject to both DPs and EPs.

Type of Discrepancy	Deceptive Patterns (DPs)	Explanatory Pitfalls (EPs)
Discrepancy between explanation and model	**Lying** John-Mathews [24], Le Merrer and Trédan [29]	**Unfaithfulness** Bove et al. [5], Jacovi and Goldberg [21], Miró-Nicolau et al. [33]
	Forging Anders et al. [1], Aïvodji et al. [2], Bordt et al. [4], Dimanov et al. [10], Lakkaraju and Bastani [27], Shahin Shamsabadi et al. [42], Slack et al. [43, 44]	**Cognitive Limitations** Freiesleben and König [16], Herman [18], Yampolskiy [54]
	Explanation Irrelevance Cabitza et al. [7], Eiband et al. [14], Watson and Floridi [52]	
Discrepancy between explanation and ground truth		**Model Imperfections** Vicente and Matute [49]
		Explanation Imperfections Bordt et al. [4], Cabitza et al. [7], Freiesleben and König [16], Luo et al. [31], Miró-Nicolau et al. [33], Morrison et al. [35], Papenmeier et al. [38, 39], Spitzer et al. [46]
Discrepancy within or between explanations	**Internal Contradictions** Fioravanti et al. [15], Kim et al. [25]	
	Unstable Explainers Bove et al. [5], Jacovi and Goldberg [21], Visani et al. [51], Zhou et al. [56]	
	Ambiguity & Disagreement Problem Bordt et al. [4], Bove et al. [5], Brughmans et al. [6], Krishna et al. [26], Müller et al. [36], Sundararajan and Najmi [47]	

by proposing a comprehensive taxonomy of quantitative EMs aligned with twelve desirable properties. From their perspective, a good explanation is one that satisfies multiple desirable properties simultaneously. The authors also emphasize that explanations often involve trade-offs between different properties, such as completeness versus correctness. Yet, this positive framing—specifying what explanations should ideally achieve—does not systematically address how different configurations of wrongness across these properties might carry distinct risk profiles or when violations might be acceptable.

A third perspective, focuses on *which aspect* of the XAI method the evaluation targets [45]. The authors propose classifying EMs based on which step of the XAI process they address. They introduce three categories corresponding to whether methods assess the quality of explanations themselves, the user's understanding of those explanations, or the satisfaction of broader societal goals such as trust and fairness. Their approach aims to provide a more comprehensive framework for evaluating XAI methods and to guide future work toward context-appropriate evaluation strategies. Nevertheless, like other evaluation perspectives, this framework does not distinguish between explanations that are merely imperfect and those that are systematically misleading.

2.3 A Wrongness-Focused Perspective of Explanations

Despite these evaluatory advances in XAI research, a fundamental challenge remains: in practice, achieving a perfect explanation (across all quality dimensions) is oftentimes even beyond reach (e.g., 46), not least because models themselves are often imperfect. When striving for a perfect explanation is costly or impossible, the focus thus shifts to trading off risks and benefits. In such cases, it is worthwhile to understand and discuss the risks and benefits of different types of wrongness. Existing frameworks guide optimization toward desirable properties or appropriate EMs but remain largely silent on how to characterize and prioritize different types of wrongness. An explanation can score well on multiple quality dimensions while still being totally wrong in some sense, or conversely, violate several quality criteria while still serving its intended purpose sufficiently.

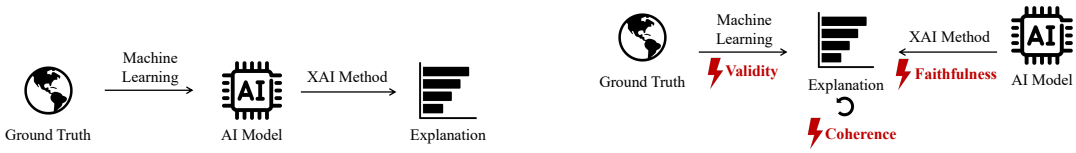
In contrast to prior contributions, our work examines explanation quality from a *wrongness-focused perspective*, motivated by insights from social sciences [32], that systematically analyzes when and how explanations are wrong. This perspective enables us to, first, distinguish between explanations that are merely imperfect and those that are systematically misleading, and second, understand the *risk factors and benefits* of different types of wrong explanations. By explicitly characterizing how explanations can be wrong across multiple dimensions, we identify distinct "shades of wrongness" that vary in their risk and benefit potential.

3 Dimensions of Wrongness

In recent years, there has been a surge in critical examinations of the limitations, pitfalls, and risks associated with XAI approaches. Conceding that explanations can be imprecise, misleading, or wrong has evolved from a minor footnote into common sense. Building on this insight, multiple attempts have been made to establish a foundational vocabulary and measurable criteria for evaluating explanation quality [12, 37]. In this work, we deliberately narrow the focus to explanatory content and, more specifically, to the question of what it means for such content to be "wrong". While the notion of explanation quality is often framed in positive terms—what explanations should ideally achieve—we argue that a systematic account of wrongness is equally necessary. Without a precise understanding of how and why explanations fail, evaluation criteria risk remaining underspecified or misleading.

One of the content-related properties proposed by Nauta et al. [37] is "correctness", which refers to the degree to which an explanation faithfully represents the behavior of the underlying model. Faithfulness in this sense has become a central criterion in the XAI literature. However, we contend that faithfulness alone is insufficient to rule out wrong explanations. An explanation may be perfectly faithful to a model while still being wrong in a more substantive sense. If a model produces blatantly erroneous outputs, and an explanation accurately reflects the internal logic leading to these outputs, the explanation remains problematic—at least with respect to the ground truth it purports to explain [46]. Conversely, as Jacovi and Goldberg [21] note, explanations that appear plausible or intuitively convincing are not necessarily guaranteed to truthfully reflect a model's internal reasoning processes. To better understand these nuances, we introduce three dimensions of wrongness: *faithfulness*, *validity*, and *coherence*. Faithfulness concerns the relationship between an explanation and the underlying model; validity captures the relationship between the explanation and the world, that is, whether the explanation aligns with relevant empirical or causal observations; coherence refers to the logical consistency of the explanation. Figure 1 depicts this explanation-centered perspective and contrasts it with a more traditional view that conceives explanations as the result of a modelling pipeline[2].

[2]We are deliberately leaving the dimensions of the data and the user out of consideration to avoid further complexity (see Section 7).



(a) Traditional pipeline view of explanations from abstracting patterns of the world to extracting explanations from an AI model.

(b) Explanation-centered view to highlight the relationship to the model (*faithfulness*), the world (*validity*), and itself (*coherence*).

Fig. 1. Comparison between traditional views on explanations (a) and our explanation-centered view to highlight wrongness (b).

Against this background, we argue that technically oriented approaches to explanatory quality in XAI should be complemented by perspectives from Science and Technology Studies (STS), philosophy of science, and epistemology. In particular, we reject accounts that treat computational systems and human cognition as subject to fundamentally different standards of explanation and evaluation. Such approaches risk obscuring the normative assumptions that shape explanatory practices across both human and technical domains. AI systems are often described as neutral instruments whose failures can be traced back to human decisions in areas such as programming or data selection. STS scholarship has challenged this separation between neutral technology and value-laden social practice [11, 23]. Technical systems are shaped by institutional priorities, social assumptions, and established forms of practice, while human action takes place within environments structured by technical infrastructures. Explanatory systems in AI therefore cannot be separated from the contexts in which they are developed and used. Explanations emerge within sociotechnical arrangements that shape both their meaning and their epistemic function. We adopt this perspective in our analysis of wrongness in XAI. We understand explanations as sociotechnical artefacts embedded in institutional and epistemic practices. Their evaluation depends not only on formal properties, but also on how they structure interpretation and guide action within specific contexts. XAI wrongness therefore cannot be understood exclusively in terms of formal accuracy or computational performance. It also concerns the conditions under which explanations become misleading, unhelpful, or even normatively problematic within concrete settings of use. These contextual dimensions are discussed in greater detail. This understanding of explanation also aligns with recent work in XAI that draws on insights from the social sciences and emphasizes the contextual and pragmatic dimensions of explanatory practice (e.g., 8, 57).

At the same time, drawing on well-established accounts from the philosophy of science, we argue that such artefacts remain subject to epistemic constraints, including the requirements of representational adequacy and explanatory relevance. Adopting a sociotechnical perspective, we consider criteria such as *Faithfulness*, *Coherence*, and *Validity* to be the minimum requirements for evaluating explanations within specific social and institutional settings.

Validity: The Relationship between Explanation and the Ground Truth. *Validity* concerns whether the explanation reliably reflects the ground truth (external world) that the model seeks to describe. A lack of validity indicates that an explanation misrepresents the actual patterns or causal relations in the ground truth—e.g., by relying on features that do not genuinely contribute to the outcome of interest. From an epistemological perspective, validity can be understood as a constraint on external adequacy: explanations are valuable only if they track the relevant aspects of the ground truth, rather than merely reproducing the model's internal logic. Invalid explanations thus fail to provide trustworthy knowledge about the world, even if they are fully faithful to the model.

Faithfulness: The Relationship between Explanation and Model. While validity evaluates how well an explanation represents the ground truth, *Faithfulness* describes how well the explanation represents the underlying AI model. A lack of faithfulness suggests that the explanation wrongly depicts the AI model we aim to scrutinize, e.g., because it uses different features than the AI model. From a philosophy of science and epistemology perspective, faithfulness can be understood as a representational constraint: explanations are required to stand in an appropriate correspondence relation to their target system. This mirrors classical concerns about scientific representation, according to which models and explanations are epistemically valuable only insofar as they adequately track the properties and relations of the ground truth they purport to describe. In this sense, unfaithful explanations do not merely reduce interpretability but fail to meet a minimal standard of epistemic adequacy [53].

Coherence: The Relationship of the Explanation to Itself. Coherence describes whether and how the explanation maintains a consistent and logical internal structure. A lack of coherence suggests that the explanation is wrong because it is logically conflicting in itself, e.g., when it provides examples that logically exclude each other [46]. From an epistemological perspective, coherence can be understood as a requirement for internal epistemic integrity: an explanation cannot be considered logically or inferentially sound if it contradicts itself. In this sense, incoherent explanations can not be valid, as they fail to provide sound knowledge about the world. At the same time, coherence is conceptually distinct from faithfulness: an explanation can remain faithful to a model even if it is incoherent, e.g., when the model is non-deterministic, or itself encodes conflicting patterns.

4 Explanation Archetypes

Table 2. Typology of explanation archetypes derived from the eight combinations of binary wrongness dimensions. The 5 shades of wrong are highlighted with grey scales.

Valid	Faithful	Coherent	Archetype	Shade of Wrong
✓	✓	✓	Perfect Explainer	(a)
✓	✗	✓	Perfect Liar	(b)
✗	✓	✓	Illuminator	(c)
✗	✓	✗	Confused Illuminator	(d)
✗	✗	✓	Storyteller	(e)
✗	✗	✗	Hallucinator	(f)
✓	✓/✗	✗	*Logically excluded*	

In the following, we will explicate different configurations of the three dimensions by simplifying them into binary values (true or false). While the dimensions are more gradual in reality, this allows us to create eight archetypes, two of which are logically excluded[3]. To make these archetypes more tangible and intuitive, we use the popular wolf-husky image classification example from Ribeiro et al. [40]. In the original example, a model is trained to accurately classify pictures of wolves and huskies without learning any meaningful characteristics about the animals and relying on background information instead. In our example, we assume we know the real features the model is using to make a prediction (i.e., the animal's shape, the background, or the moon). From that, we can iterate through the explanations archetypes (see Figure 2).

[3]The two remaining combinations (Valid-Faithful-Incoherent and Valid-Unfaithful-Incoherent) are deliberately excluded because we are not aware of any realistic examples. An explanation can be internally logical (coherent) and describe valid relations, or it can be internally illogical (incoherent) and describe invalid relations. However, an explanation cannot be internally incoherent and accurately describe valid relations at the same time; that would require assuming valid structures in the world can be described incoherently, which, to the best of our knowledge, does not occur.

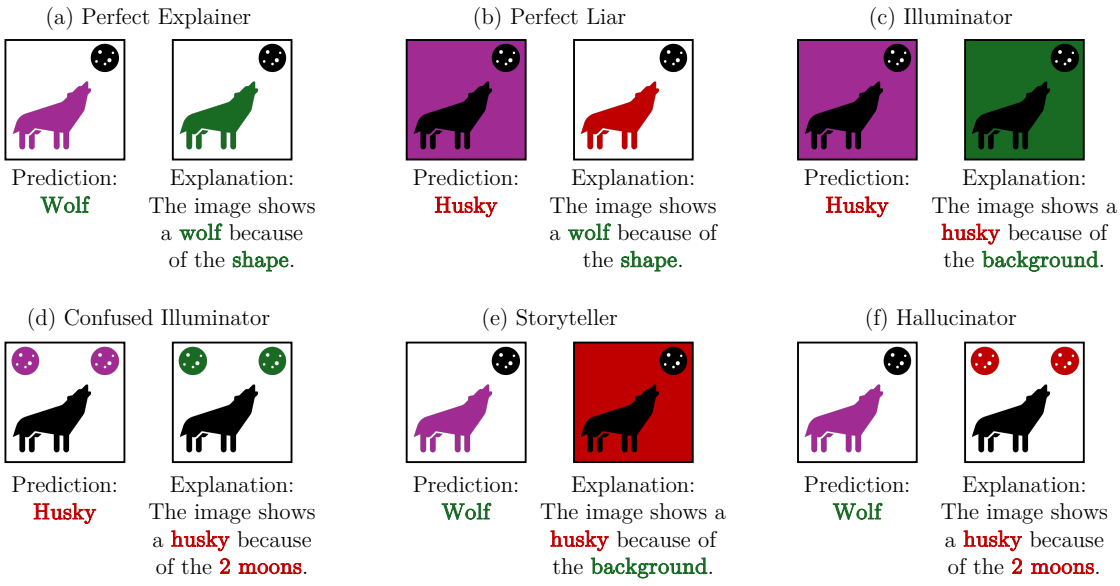


Fig. 2. Visual examples of the explanation archetypes. Each subfigure presents the model prediction with the actually used features highlighted in purple on the left and the corresponding explanation on the right. The explanation includes a feature heatmap (green if faithful, red if unfaithful), a textual description of the prediction (green if valid, red if invalid), and a rationale (green if coherent, red if incoherent).

The Perfect Explainer: Valid–Faithful–Coherent. As a perfectly correct baseline, we assume an explanation that is faithful, valid, and coherent, meaning that it identifies the relevant factors contributing to the outcome and accurately reflects how the model utilized those factors. This archetype represents the gold standard for explanations. In the example shown in Figure 2 (a), both the predictive model and the heatmap explanation focus on the image region corresponding to the wolf.

The Perfect Liar: Valid–Unfaithful–Coherent. In this case, the explanation appears plausible and reflects the ground truth; however, it fails to accurately represent the underlying model. This discrepancy is particularly problematic because the explanation is coherent and describes meaningful relations, while neglecting its primary purpose: faithfully explaining the model's reasoning. In the example shown in Figure 2 (b), the model attends to the image background, whereas the heatmap explanation focuses on the animal's shape—resulting in an explanation that is valid and coherent but ultimately unfaithful.

The Illuminator: Invalid–Faithful–Coherent. In this case, the explanation is faithful to the AI model and appears plausible, yet it leads to invalid conclusions. Such an explanation serves the traditional purpose of model debugging by revealing that the model is not operating as intended. In the example shown in Figure 2 (c), both the model and the explanation focus on the background, making the explanation faithful but invalid.

The Confused Illuminator: Invalid–Faithful–Incoherent. In this case, the explanation faithfully reflects an invalid and incoherent model. Because the model itself—or parts of it (e.g., when processing out-of-distribution samples)—must exhibit incoherence, this scenario is likely rare. While such explanations can reveal flaws in the AI model (e.g., inadequate

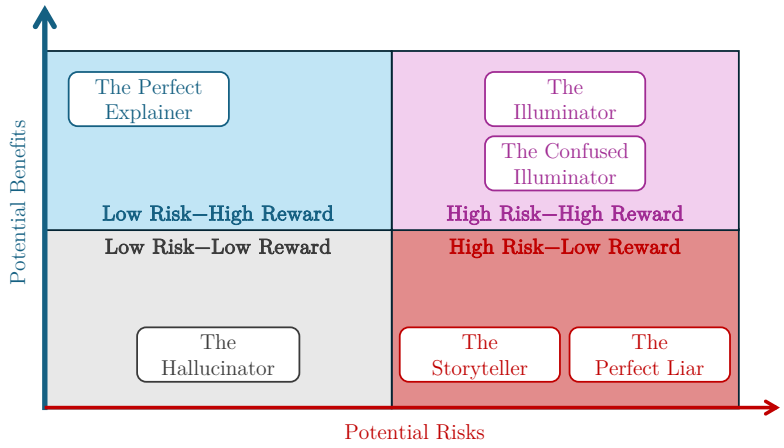


Fig. 3. The explanation archetypes are positioned according to their potential risks and benefits, which creates two dimensions and four quadrants. The bottom-left quadrant represents neutral explanations that neither pose significant risks nor benefits. The top-left quadrant represents explanations that fulfill their purpose without significant risks of misleading interpretations. Explanations in the top-right quadrant can expose flaws of the underlying AI model, but also pose the risk of accepting these flaws, commonly associated with *EPs*. The bottom-right quadrant represents explanations that carry the maximum risk of deception and offer little added value; these are strongly associated with *DPs*.

feature engineering), they may be difficult to interpret due to their inherent incoherence. In the example shown in Figure 2 (d), both the model and the heatmap explanation highlight patterns of two moons, making the explanation faithful yet invalid and incoherent (because there can only be one moon in an image).

The Storyteller: Invalid–Unfaithful–Coherent. In this case, the explanation is coherent but neither valid nor faithful to the model. In other words, it presents a plausible narrative that is revealed as invalid only when the valid relationships are known or faithfulness is assessed. In the example shown in Figure 2 (e), the model attends to the animal's shape, whereas the heatmap explanation focuses on the background—resulting in an explanation that is coherent yet invalid and unfaithful to the AI model.

The Hallucinator: Invalid–Unfaithful–Incoherent. In this case, the explanation is wrong in every sense of the word—neither faithful to the AI model, nor valid, nor coherent. Such an explanation is difficult to interpret and is likely to reveal itself as neither meaningful nor trustworthy. In the example shown in Figure 2 (f), the heatmap explanation highlights two moons, which is not only incoherent in itself but also unfaithful to the model and invalid.

5 Risk and Benefit Analysis of Archetypes

Our systematic characterization of wrongness dimensions and explanation archetypes allows us to derive general insights about their unique potential risks and benefits. Most importantly, we observe that the risk of wrongness is not strictly additive; i.e., being wrong in one dimension can pose a greater risk than being wrong in other (or all) dimensions, for example, because the wrongness is more difficult to detect. Figure 3 visualizes the archetypes by assigning them to one of four benefit-risk quadrants.

However, it is important to note that for practical insights, our wrongness dimensions should not be treated as binary variables but rather as a spectrum. Also, in practice, the assessment of any archetype's risks and benefits is highly context-dependent. Factors such as the application domain, the stakes involved, the affected stakeholders, and institutional norms all shape whether a given type of wrongness is harmful or acceptable [57]. This aligns with insights from the embedded ethics and social sciences approach, which emphasizes that ethical evaluation of AI systems cannot be meaningfully conducted in the abstract but must account for the specific sociotechnical setting of deployment [48].

To this end, we will provide a running example of a hypothetical recruiting system, a high-risk AI system that is used to screen and recommend applicants for a job offering. In our scenario, the recruiting system provides feature-based explanations with relevant characteristics from the applicants' CVs and is overseen by a human-in-the-loop who sends out job interviewes based on the recommendations and uses the explanations to formulate rationales for the rejected applicants. When we discuss risks, we explicitly include risks that occur during model development (e.g., a false sense of security) as well as downstream risks through human-AI interaction (e.g., overreliance and underreliance).

The *low risk–high reward* quadrant represents helpful decision support, where users know what to expect, as decisions are based on valid factors and are faithfully explained. Obviously, this is where the *Perfect Explainer* is placed. Unfortunately, this idealized scenario is rare and difficult to ensure. In the example of the recruiting system, we would need causal or ongoing productive evaluation to ensure that it correctly identifies relevant qualifications. Additionally, we would need to design and evaluate an XAI method that remains faithful across samples and over time. In this scenario, applicants and hiring managers could easily verify and, if necessary, contest the recommendations.

The *low risk–low reward* quadrant represents neutral explanations that provide no significant benefit, but are also so irritatingly wrong that there is no substantial risk of misplaced trust. This is where the *Hallucinator* is placed. Because the explanations are logically inconsistent and not causally grounded in reality, the lack of faithfulness bears low potential risk, as we can expect developers and users to easily recognize that these explanations are not reliable and should not be trusted. Still, a residual risk remains that wrong AI advice may be adopted despite these warning signals. In the example of the recruiting system, a *Hallucinator* explanation could reference to irrelevant excerpts from the CV and provide contradicting rationales (e.g., "too young" and "too old"). While such a system can theoretically lead to ineffective hiring decisions or even discriminatory outcomes, the contradictions indicate severe technical flaws in the XAI method and should likely be detected by the human-in-the-loop or ideally even during development. Especially for high-risk systems, companies are incentivized to rule out such major flaws before deploying a recruiting system.

In the *high risk–high reward* quadrant, explanations have the valuable potential to expose flaws of the underlying AI model. However, they also pose the risk of erroneously accepting these flaws, leading to *EPs* (e.g., 25). In this quadrant, we place explanation archetypes that can be used for model debugging when detected during model development. Coupled with faithfulness evaluation and ground truth labels, the *Illuminator* can effectively call out flawed predictions, which makes it perfectly suitable for model debugging during development. It sits slightly above the *Confused Illuminator*, whose incoherent presentation impedes understanding the model's reasoning. In the example of the recruiting system, an *Illuminator* explanation could faithfully reveal an applicant's zip code—a proxy feature for ethnicity—as an invalid driving factor for the recommendation. During development and auditing, this could be a highly valuable observation that can be used to rule out such undesired correlative patterns. However, without critical scrutiny, the human-in-the-loop may accept such recommendations and explanations when the explanation appears plausible and intuitive.

Lastly, the *high risk–low benefit* quadrant represents explanations that carry the maximum risk of deception without offering any epistemic benefit. These kinds of explanations usually require malicious intent and are strongly associated with *DPs* (e.g., 2). Unlike the *Hallucinator*, the *Storyteller* presents unfaithful and invalid explanations in a coherent

manner. While the explanation provides no benefit, the potential risk increases due to its apparent plausibility. The coherent presentation makes this character particularly susceptible to misuse as a DP. *The Perfect Liar* poses the highest risk due to its strong plausibility without any anchoring in the underlying AI model. This characteristic makes the Perfect Liar particularly attractive for exploitation. The critical danger lies in the fact that these causally sound statements have no connection to the underlying model's actual reasoning, making them exceptionally difficult to assess. In the example of the recruiting system, the system could include a range of discriminatory disadvantages but superficially present *Perfect Liar* explanations emphasizing relevant and legitimate features (e.g., qualifications, experience) while the actual decision was driven by gender. We can imagine cases in which companies deliberately employ such deceptive explainers to mask unacceptable biases and project a facade of objectivity that can hardly be detected by applicants without model access [29]. This presents a classic case of "fairwashing" [2] where the "right to explanation" envisioned by the EU AI Act is undermined by plausible but unfaithful explanations [20].

6 Strategies to Deal With Wrongness

Sources of wrongness in XAI are manifold. From an STS perspective, valid relationships in the world should be understood as situated [17], while at the same time they may be highly complex and difficult to learn; sometimes the AI model is poorly designed, or maybe the selected XAI method is not able to faithfully depict the AI model. In addition, we understand wrongness as a socio-technical phenomenon: wrongness is therefore not merely a technical flaw but also socially mediated. What counts as "wrong" may depend on the interpretive frameworks, expectations, and trust relations of the users engaging with the explanation. Our risk and benefits analysis of the explanation archetypes shows that each archetype requires tailored strategies to mitigate these risks. Crucially, these strategies are context-dependent and shaped not only by the technical properties of the model but also by the social, organizational, and practical conditions in which explanations are deployed. Stakeholders, institutional norms, and the associated stakes influence both the perception and the consequences of errors. In the following, we sketch strategies to address wrongness in the many cases where *Perfect Explainer* is not achievable. Taken together, risks of both DPs and EPs can be mitigated through heightened awareness of the limitations of XAI among developers, users, and affected parties.

6.1 Transparency: From Narrow to Broad XAI

When wrongness is unavoidable, an obvious strategy is to make it transparent. Typologies and criteria—such as our archetypes or the Co-12 Properties [37]—can provide important vocabularies and measures to report these limitations. For instance, EMs like infidelity [55] allow quantification of the faithfulness dimension. Establishing more benchmarks and transparency could mitigate the potential risks associated with the *Hallucinator*, *Storyteller*, and *Perfect Liar* by helping developers to detect flaws in the AI model and helping users to calibrate when to rely on the explanation (and when not). While this type of transparency is usually not at the center of XAI research, several scholars have called for a broader understanding of "XAI" [8, 9] that incorporates background information about the model and its limitations. One advantage of this broader understanding is that many of these approaches are unaffected by the dimensions of wrongness outlined in our paper. For example, the validity of model cards [34] is only dependent on the sincerity of model developers, and supplementary information like uncertainty quantification [3] or fairness EMs [8] provides crucial information to users beyond traditional explanations.

6.2 Training: Awareness Through (X)AI Literacy

If the wrongness of an explanation does not become clear through obvious incoherence (see, e.g., the *Hallucinator*), the (X)AI system may conceal its wrongness. Therefore, it becomes a crucial skill for users to critically judge whether the explanation is wrong or not. This capability begins with the awareness that an explanation can be (intentionally or unintentionally) wrong. Ziethmann et al. [57] expand the definition of AI Literacy [30] and define *XAI Literacy* "as a set of competencies that enables individuals to critically evaluate the output of XAI methods; and thus enabling AI Literacy." It is important to recognize that each application domain brings its own limitations for explanations, which may vary in severity [57]. They further highlight key aspects affected by the limitations, such as inherent and unavoidable data bias, the distinction between correlation and causation, or the discriminatory effects of parameters. To ensure *XAI literacy*, users need to be educated about the context-specific limitations and possibilities of XAI. Suitable formats could include user workshops where users are confronted with wrong explanations, informed about the types of wrongness, and learn to critically reflect on (X)AI system outputs.

6.3 Wrong for the Right Reasons

As a final remark, we want to highlight that wrongness is not *always* necessarily a negative property of explanations. As we have shown in Section 5, some archetypes pose higher risks than others, and in some contexts, certain archetypes may be acceptable despite being wrong to some extent. For example, in applications where validity is not necessary and correlation is sufficient (e.g., anomaly detection for credit card fraud), the *Illuminator* may provide perfectly useful explanations. When the underlying AI model is flawed, the *Illuminator* or the *Confused Illuminator* may even be desirable *exactly because* they are wrong. Thus, our explanation-focused view prompts the XAI community not to mitigate wrongness per se, but to be explicit about the purpose and the limitations of the employed XAI approach.

7 Conclusion and Outlook

In this paper, we presented a typology of wrongness for explanations of AI models, establishing five archetypal "shades of wrongness" that categorize the specific ways in which explanations can be wrong. With that, we move beyond a simple right/wrong dichotomy and empower developers and users to assess and communicate the quality of XAI systems more rigorously. A critical finding from our analysis is that the risks associated with these dimensions are not additive. While one might intuitively assume that an explanation failing in all three dimensions (the *Hallucinator*) poses the greatest threat, we argue that such explanations are, paradoxically, less risky because their absurdity is often self-revealing. Instead, we warn practitioners to scrutinize more subtle archetypes, such as the *Perfect Liar*, as these are significantly more misleading and susceptible to exploitation of DPs. Making these dimensions explicit is essential for advancing *XAI Literacy*.

To provide a clear conceptual foundation, our work made several simplifying assumptions. First, for simplicity, we deliberately omitted sub-dimensions of faithfulness and validity, which can both be decomposed into procedural and outcome sub-dimensions. For example, an explanation can faithfully replicate the output of the underlying model but still use different features [26]. This is an important consideration, as many popular XAI methods such as LIME [40] rely on surrogate models or output computations to approximate information about the process. Similarly, models and XAI methods can generate output that is consistent with the ground truth without picking up any meaningful causal or even correlative relationships (e.g., through spurious correlations or pure chance). The distinction between process and outcome perspectives in XAI is rarely straightforward and touches upon complex debates regarding causality in AI [8],

which we leave for future work. Second, we illustrated our archetypes primarily through heatmap explanations. Future work should test the generalizability of our typology by applying it to other XAI methods—such as example-based or counterfactual explanations—and by studying these archetypes in real-world deployment scenarios.

Moreover, future work should investigate dimensions of wrongness that lie beyond the immediate scope of explanation generation. Upstream, this includes the integrity and suitability of the data [22]. Downstream, the analysis may extend to the human element including erroneous interpretation or inappropriate behavior [46]. Lastly, explanations can be *morally wrong*, raising the question of ethical or societal desirability in explanations. For example, even a valid explanation may violate social and ethical norms when it identifies a sensitive feature as the decisive factor for rejecting an applicant in a hiring decision. This illustrates that technical correctness does not always equate to normative alignment.

Finally, future work should empirically investigate the existence of our archetypes in real-world applications. While we provide arguments and illustrative examples for the potential risks and benefits of different archetypes, only empirical studies will demonstrate how risks and benefitis materialize in real-world settings where concrete harms can be studied. In future empirical studies, we also plan to investigate how targeted XAI literacy interventions for developers and affected stakeholders can increase productive use of imperfect explanations and mitigate the potential harms we outlined in Section 5.

As AI systems increasingly permeate high-stakes decision-making, the pursuit of "perfect" explanations is often unrealistic. By acknowledging that models and explainers are inherently imperfect, we can shift our focus from chasing an unattainable ideal to realistically managing the risks of different types of wrongness. Our typology serves as a foundation for this shift, helping practitioners prioritize the most harmful forms of wrongness to prevent misuse, foster appropriate trust, and protect vulnerable user groups from the pitfalls of misleading explanations.

References

[1] Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, et al. 2020. Fairwashing explanations with off-manifold detergent. *ICML* (Jan. 2020), 314–323.

[2] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, et al. 2019. Fairwashing: The risk of rationalization. In *Proceedings of the 36th ICML*. 161–170.

[3] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, et al. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 401–413.

[4] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike Luxburg. 2022. Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In *2022 ACM FAccT*. Association for Computing Machinery, New York,NY,United States, 891–905.

[5] Clara Bove, Thibault Laugel, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2025. Why do explanations fail? A typology and discussion on failures in XAI. arXiv:2405.13474 [cs].

[6] Dieter Brughmans, Lissa Melis, and David Martens. 2023. *Disagreement amongst counterfactual explanations: How transparency can be deceptive*. Technical Report.

[7] Federico Cabitza, Caterina Fregosi, Andrea Campagner, and Chiara Natali. 2024. Explanations Considered Harmful: The Impact of Misleading Explanations on Accuracy in Hybrid Human-AI Decision Making. Springer Nature Switzerland, Cham, 255–269.

[8] Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. A Critical Survey on Fairness Benefits of Explainable AI. In *The 2024 ACM FAccT*. ACM, New York, NY, USA.

[9] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, et al. 2021. Who needs to know what, when?: Broadening the Explainable {AI (XAI)} Design Space by Looking at Explanations Across the {AI} Lifecycle. In *DIS Conference 2021*. 1591–1602.

[10] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. 2020. You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods. In *SafeAI@AAAI 2020*.

[11] Roel Dobbe and Anouk Wolters. 2024. Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context. *Minds and Machines* 34, 1 (2024), 1–28.

[12] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [cs, stat].

[13] Upol Ehsan and Mark Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5 (June 2024), 100971.

[14] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI*. Association for Computing Machinery.

[15] Stefano Fioravanti, Francesco Giannini, Pietro Barbiero, Paolo Frazzetto, Roberto Confalonieri, et al. 2025. Categorical Explaining Functors: Ensuring Coherence in Logical Explanations. *Proceedings of the KR* 22, 1 (2025), 306–315.

[16] Timo Freiesleben and Gunnar König. 2023. Dear XAI Community, We Need to Talk! In *Explainable Artificial Intelligence: First World Conference, xAI 2023, Lisbon, Portugal, July 26–28, 2023, Proceedings, Part I*. Communications in Computer and Information Science, Vol. 1901.

[17] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies* 14, 3 (1988), 575–599.

[18] Bernease Herman. 2017. The Promise and Peril of Human Evaluation for Model Interpretability. In *31st NIPS Conference 2017)*.

[19] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects.

[20] Anton Hummel, Håkan Burden, Susanne Stenberg, Jan-Philipp Steghöfer, and Niklas Kühl. 2026. The Bidirectional Relationship Between XAI and Regulation: Operationalizing XAI for the AI Act.

[21] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, 4198–4205.

[22] Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, et al. 2022. *Data-Centric Artificial Intelligence*. Technical Report.

[23] Sheila Jasanoff. 2004. *States of Knowledge: The Co-Production of Science and Social Order*. Routledge.

[24] Jean-Marie John-Mathews. 2022. Some critical and ethical perspectives on the empirical turn of {AI} interpretability. *Technological Forecasting and Social Change* 174 (Jan. 2022), 1–29. Number Of Volumes: 121209.

[25] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, et al. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies *(CHI '25)*.

[26] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, et al. 2025. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. doi:10.48550/arXiv.2202.01602

[27] Himabindu Lakkaraju and Osbert Bastani. 2020. {"How do I fool you?" Manipulating} User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[28] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, et al. 2021. What Do We Want From Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296 (July 2021), 103473. arXiv:2102.07817 [cs].

[29] Erwan Le Merrer and Gilles Trédan. 2020. Remote explainability faces the bouncer problem. *Nature Machine Intelligence* 2, 9 (Jan. 2020), 529–539.

[30] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16.

[31] Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Zhu Xiaodan. 2022. Evaluating Explanation Correctness in Legal Decision Making.

[32] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38.

[33] Miquel Miró-Nicolau, Antoni Jaume-i Capó, and Gabriel Moyà-Alcover. 2024. Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets. *Artificial Intelligence* 335 (Oct. 2024), 104179.

[34] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, et al. 2019. Model Cards for Model Reporting. In *Proceedings of the FAccT*. ACM, New York, NY, 220–229.

[35] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, et al. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–39.

[36] Sebastian Müller, Vanessa Toborek, Katharina Beckh, et al. 2023. An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, Danai Koutra et al. (Eds.). Vol. 14171. Cham, 462–478.

[37] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, et al. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *Comput. Surveys* 55, 13s (Dec. 2023), 1–42.

[38] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust in {AI}. In *Proceedings of the IJCAI 2019*. IJCAI, 94–100.

[39] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in {AI}. *ACM TOCHI* 29, 4 (Jan. 2022), 1–33.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. {"Why should I trust you?" Explaining the predictions of any classifier}. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

[41] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2024. On Explanations, Fairness, and Appropriate Reliance in Human-{AI} Decision-Making. *ACM CHI 2024* (Jan. 2024).

[42] Ali Shahin Shamsabadi, Mohammad Yaghini, Natalie Dullerud, Sierra Wyllie, Ulrich Aïvodji, et al. 2022. Washing The Unwashable : On The (Im)possibility of Fairwashing Detection. *Advances in Neural Information Processing Systems* 35 (Jan. 2022), 14170–14182.

[43] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.). Curran Associates, Inc, 6275.

[44] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, et al. 2020. Fooling {LIME} and {SHAP}: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[45] Timo Speith and Markus Langer. 2023. A New Perspective on Evaluation Methods for Explainable Artificial Intelligence (XAI). In *2023 IEEE 31st REW*. IEEE, Hannover, Germany, 325–331.

[46] Philipp Spitzer, Katelyn Morrison, Violet Turri, Michelle Feng, et al. 2025. Imperfections of XAI: Phenomena Influencing AI-Assisted Decision-Making. *ACM Trans. Interact. Intell. Syst.* 15, 3 (Sept. 2025), 17:1–17:40.

[47] Mukund Sundararajan and Amir Najmi. 2020. The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 9269–9278. https://proceedings.mlr.press/v119/sundararajan20b.html

[48] Daniel W. Tigard, Maximilian Braun, Svenja Breuer, Amelia Fiske, Stuart McLennan, and Alena Buyx. 2024. Embedded Ethics and the "Soft Impacts" of Technology. *Bulletin of Science, Technology & Society* 44, 3-4 (Dec. 2024), 73–83. doi:10.1177/02704676241298162

[49] Lucía Vicente and Helena Matute. 2023. Humans inherit artificial intelligence biases. *Scientific Reports* 13, 1 (Jan. 2023), 15737.

[50] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (Dec. 2021), 89–106.

[51] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. 2022. Statistical Stability Indices for LIME: Obtaining Reliable Explanations for Machine Learning Models. *Journal of the Operational Research Society* 73, 1 (2022), 91–101. doi:10.1080/01605682.2020.1865846

[52] David S. Watson and Luciano Floridi. 2021. The Explanation Game: A Formal Framework for Interpretable Machine Learning. In *Ethics, Governance, and Policies in Artificial Intelligence*. Vol. 144.

[53] Michael Weisberg. 2013. *Simulation and similarity: Using models to understand the world.* Oxford University Press, New York, US.

[54] Roman V. Yampolskiy. 2020. Unexplainability and Incomprehensibility of AI. *Journal of Artificial Intelligence and Consciousness* 07, 02 (Sept. 2020), 277–291.

[55] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, et al. 2019. On the (In)fidelity and Sensitivity of Explanations. In *Advances in NeurIPS*, Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html

[56] Zhengze Zhou, Giles Hooker, and Fei Wang. 2021. S-LIME: Stabilized-LIME for Model Explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. Association for Computing Machinery, 2429–2438. doi:10.1145/3447548.3467274

[57] Paula Ziethmann, Anton Hummel, Alexander Althammer, Kerstin Schlögl-Flierl, et al. 2025. From Embedded Ethics to Explainable AI: Advancing Interdisciplinary Collaboration in Medical AI Development.