Diagnosing AI Explanation Methods with
Folk Concepts of Behavior

Alon Jacovi   ALONJACOVI@GMAIL.COM
Bar Ilan University

Jasmijn Bastings   BASTINGS@GOOGLE.COM
Google DeepMind

Sebastian Gehrmann   SGEHRMANN8@BLOOMBERG.NET
Google Research

Yoav Goldberg   YOAV.GOLDBERG@GMAIL.COM
Bar Ilan University,
Allen Institute for Artificial Intelligence

Katja Filippova   KATJAF@GOOGLE.COM
Google DeepMind

Abstract

We investigate a formalism for the conditions of a successful explanation of AI. We consider "success" to depend not only on what information the explanation contains, but also on what information the human explainee understands from it. Theory of mind literature discusses the folk concepts that humans use to understand and generalize behavior. We posit that folk concepts of behavior provide us with a "language" that humans understand behavior with. We use these folk concepts as a framework of *social attribution* by the human explainee—the information constructs that humans are likely to comprehend from explanations—by introducing a blueprint for an explanatory narrative (Figure 1) that explains AI behavior with these constructs. We then demonstrate that many XAI methods today can be mapped to folk concepts of behavior in a qualitative evaluation. This allows us to uncover their failure modes that prevent current methods from explaining successfully—i.e., the information constructs that are missing for any given XAI method, and whose inclusion can decrease the likelihood of misunderstanding AI behavior.

1. Introduction

In the development of explanation methods of AI systems, there is often a strong focus on the side of the explainer, but little attention is being paid to the exchange of information between the explainer and the explainee (Carvalho et al., 2019; Sokol & Flach, 2020; Islam et al., 2021). In particular, when explanation methods are introduced, they are typically motivated by being able to satisfy certain mathematical properties, which are not necessarily grounded in the needs of the explainee (Miller, 2019; Rutjes et al., 2019). Yet, explainees have different experiences and expertise and may thus not understand an explanation in the intended way. We aim to formalize what explainees may understand about AI processes as a result of explanations, and how this understanding may differ from what the explanation attempted to communicate. We refer to the information which the explainee comprehends as the explainee's mental model. In this work, we are concerned with AI explanations
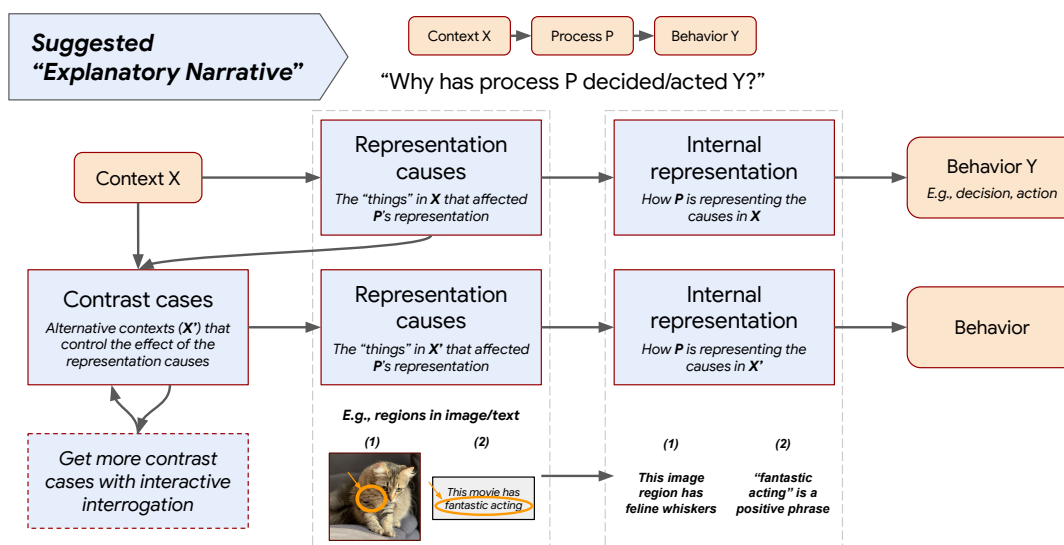
Figure 1: Schematic of an explanatory narrative, as explored in this work. The narrative communicates a causal chain composed of two categories of causes: The objective causes in the context, and the actors' subjective interpretation of those causes. The causes' role is communicated against alternative contexts (contrast cases) that intervene on them. This paper develops the narrative structure, justifies it with precedence, and applies it to modern XAI methodologies to derive useful insights about the path towards successfully explaining complex AI.

as *explanations describing AI processes* ("why did this behavior manifest?"), rather than explanations as justifications to AI decisions ("why is this behavior correct?") (Kakas & Michael, 2020; Kim et al., 2021).

XAI methods can fulfill one or more desiderata for what explanations "should" satisfy: For example, that explanations should accurately describe the AI system they are explaining (Gilpin et al., 2018; Rudin, 2018; Lakkaraju et al., 2019); or be sufficient, so that no crucial information is missing (Yu et al., 2019; Linardatos et al., 2021); or be minimal, so that no redundant information is given (Lei et al., 2016; Linardatos et al., 2021); and so on (Lipton, 2018). Such constraints are given mathematical form, and then argued for by demonstrating that XAI methods which do *not* uphold the mathematical constraints, fail in some core utility (Alvarez-Melis & Jaakkola, 2018; Feng et al., 2018; Baan et al., 2019).

This method of selecting desiderata is appealing in its flexibility, but without knowing the cognitive principles behind such desiderata, which are often born from AI practitioners' intuitions, we cannot say for certain whether, or *why*, they are necessary or useful properties of explanations that successfully communicate information.[1]

---

1. While such questions can be answered through user studies and a better understanding of user experiences and mental models, researchers often put explanations in the hands of unknown users through the release of tools (e.g., Tenney et al., 2020; Kokhlikyan et al., 2020) without knowing how their users will interpret the results, regardless of the axioms that are being satisfied. Moreover, even when user

In section 2 we address this problem by characterizing "successful explanations" in a consistent framework, rather than via a set of axioms (e.g., faithfulness or sufficiency), by pivoting the root of the analysis from the formal properties of the explainer to the cognitive properties of the explainee. The set of desiderata of a successful explanation is then *derived from* this foundational framework, as desiderata that enable a specific formalism of success—coherence—rather than being treated as axioms. To develop our framework, we draw inspiration from psychological and philosophical research in the field of *theory-theory* (Morton, 1980): The study of how humans model the outside world for the purpose of generalizing, understanding and explaining phenomena.

Research in this area additionally points to *biases*, or *habits*, that a human explainee commonly exhibits when leveraging prior knowledge in comprehending explanations of behavior. One of these habits is to understand non-human processes by drawing analogies to human behavior (Section 3.1). This makes the area of *theory of mind* and *folk psychology*, the study of how humans model human behavior, relevant to characterize humans' understanding of AI explanations and AI behavior (Miller, 2019). In Section 3.2 we investigate how folk concepts of behavior can help us to construct causal explanations of behavior.

We apply this framework to the XAI literature (Section 4), and find that many XAI mechanisms can be aligned with these folk concepts of behavior. This lets us identify what different XAI methods are *missing* which could potentially increase their ability to communicate information successfully. We do this by observing what the explanation method fails to communicate among the causal explanatory narrative in Figure 1. The missing component is in danger of being incorrectly speculated upon by the explainee, which can cause the explainee to misunderstand the explanation.

Below we summarize our primary findings:

1. We say that behavior has been successfully explained if the explainee's mental model is *coherent*, in that no contradictions are found between it and additional observations of behavior (Section 2).

2. An explainee's understanding of behavior can be conceptualized with multiple components: The internal representation of the behaving actor(s), the things that affected this representation, and the things that affected the outcome without affecting the representation. If any components are missing, the explainee may incorrectly assume them, leading to a misunderstanding of the explained process (Section 3).

3. We show that for a wide variety of current explanation methods, each of them fails the completeness test—i.e., is missing at least one of the components that compose the explanatory narrative (Section 4).

Based on these findings, we conclude that to minimize the possibility of misunderstanding by human explainees' in XAI, there are two tools at our disposal (Section 5): (1) Communicating information via a "complete" causal narrative; (2) Using interactivity to explore contrast cases that intervene on the narrative's suggested causes, as a medium of resolving contradictions methodologically.

---

studies are conducted, studying explanations in isolation is not a replacement for studying them after their deployment in actual systems (Bucinca et al., 2020).

2. Defining Successful Explanations

When comprehending an explanation to some event, the human explainee establishes a hypothesis about the event's history. We refer to this as the explainee's *mental model* (Payne, 2003). Viewing the explanation as a "function" (Lombrozo, 2006), the mental model can be considered as the outcome of the explanation. The conditions to a successful explanation are therefore conditions about the explainee's mental model (Sreedharan et al., 2021).

2.1 Coherent Mental Models

The cognitive science literature[2] often describes the goal of an explanation for the explainee as *generalization and prediction* (Woodward & Hitchcock, 2003; Lombrozo, 2006; Williams & Lombrozo, 2010; Bradley, 2017). This means that an explainee develops a "coherent" hypothesis about the circumstances that led to the explained event which is consistent even for new events (Murphy & Medin, 1985; Johnson-Laird & Byrne, 2002), and enables them to make predictions about these events (also known as *explanatory unification,* (Kitcher, 1981) and *consilience* (Thagard, 1988)). In the case of AI, this means generalizing to other instances of AI behavior. Therefore, in this work we consider an explanation as "successful" if it produces a mental model which is coherent across instances of AI behavior.

The principal constraint posed by coherency is that there are no contradictions between the explainee's mental model and additional observations (contrast cases). For example, when hiding a ball under a cup, the theory that the ball continues to exist is consistent with (does not contradict) the new observation of the ball when removing the cup. This insight relies on the explainee's mental model of object permanence.[3]

2.2 Implications

The definition of successful explanation as a function of coherent mental models implies several interesting conclusions:

Explanation "correctness" is not explicitly part of this definition. A recent trend of the XAI evaluation literature pertains to the *correctness of the explanation with respect to the AI*: Whether an explanation faithfully represents information about the model. The literature in this area establishes that XAI methods, as lossy approximations of the AI's reasoning process, are not completely faithful (Adebayo et al., 2018; Ghorbani et al., 2019) and that completely faithful and human-readable explanations are likely an unreasonable goal (Jacovi & Goldberg, 2020). Various relaxed measures of faithfulness were proposed (Appendix A).

However, human-to-human explanations also often do not provide correctness guarantees and yet are common and accepted. While an explanation should not "incorrectly" describe the event history, some allowance is permitted on the uncertainty of whether the explanation

---

2. We focus on the cognitive and developmental function of explanation in humans, as opposed to the social utility of explanation, as the uses of explanation in society (e.g., teaching, assigning blame) build on this core cognitive function.

3. In the more complex context of AI, a similar understanding can be facilitated through training programs and instructional aid that shape mental models of humans about AI behavior (Hanisch et al., 1991; Gehrmann et al., 2020).

is considered correct, in the absence of ground truth. This allowance manifests by using *coherence*, rather than *correctness*, due to this intractability. In Appendix A we discuss how empirical quality measures of AI explanations, developed in recent years by the XAI community, also in fact capture coherence.

In this perspective, correctness or faithfulness can be considered to be a useful *but not necessary* condition to successful explanations—since faithfulness can contribute to coherence, but is not the only means of doing so.

Coherence is characterized by an *empirical* budget allotted to proving or refuting it. Coherence positions the success metric of an explanation as an empirical measure rather than a theoretical one. If no contradiction is found after a "sufficient enough" search, an explanation is deemed "correct enough" (Sellars, 1963; Kitcher, 1981; Lehrer, 1990; Mayes, 2022).[4]

Explanation is interactive: Lack of coherence—the existence of contradictions— is *not* a failure state. The explainee establishes a mental model as a result of explanation via an *iterative* process, rather than one-time. This means that if coherence was refuted, i.e. contradictions arise, the mental model is deemed insufficient and can be *adjusted* by the explainee into one for which the contradiction is resolved (Shvo et al., 2022). This process, if repeated until no contradictions are found, results in a coherent mental model, and the entire process is designated as explanation. Since each step in the process is conditioned on explainee's current mental model and the contradictions that are observed by the previous iteration—explanation in its ideal form is *interactive* (Strobelt et al., 2018; Miller, 2019; Gehrmann et al., 2020; Kirchler et al., 2021).

3. Composition of an Explanatory Narrative

Assume that our goal now is to enable explainees to establish coherent mental models of an AI's behavior, according to the definition in Section 2. In this section we look at precedence in theory of mind research (Section 3.1) to construct an explanatory narrative that can aid this goal (Section 3.2).

3.1 Anthropomorphic Bias and Perceived Intentionality

Intentionality is a central concept in models of folk theory of mind (Karniol, 1978; Knobe & Malle, 1997; Burra & Knobe, 2006): It refers to the power of mind to internally represent things about the world.[5] When we comprehend explanations about events, we intuitively do so with respect to "actors" which hold internal representations, and whose behaviors had a causal role on the event (Figure 2).

The explainee's assumptions about explained events are potentially biased with respect to how humans think and behave: If there is an actor in the event's history, we may

---

4. Precise affordances of this budget is beyond our scope, and should be considered societal or regulatory in nature. For use cases with large state spaces, e.g., language generation or reinforcement learning, the problem of summarizing agent behavior under a constrained budget has been studied (e.g., Amir et al., 2019).

5. Note that this sense of the word "intentionality" in philosophy differs from the colloquial sense of intent in English.
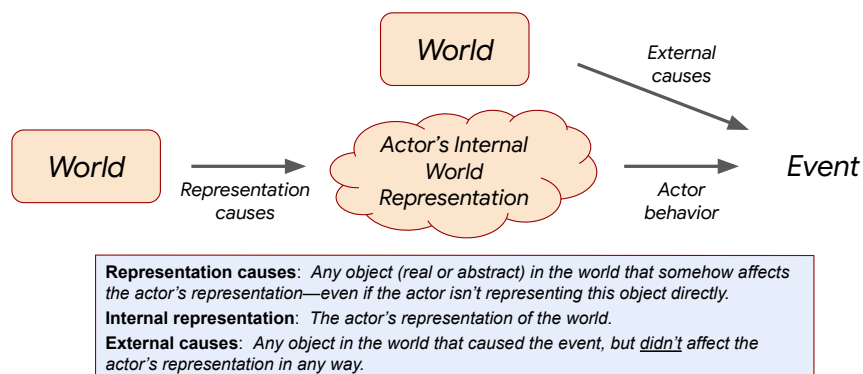
Figure 2: Folk concepts of behavior (adapted from Malle (2003)). Research shows that humans understand and explain events along these concepts. See Section 3.2 for description and examples.

potentially understand this actor (human or not) by imagining how we may have acted in the actor's circumstances, implicitly assigning a mental representation to the actor (Culley & Madhavan, 2013).

When the actor is not human, we refer to this as *anthropomorphic bias*. This bias is widespread and common (Dacey, 2017; Johnson, 2018). For example, Heider and Simmel (1944) found that humans attribute human-like behavior to simple moving shapes. Regardless of the nature of extent of this bias, if the explainee can view the AI as an actor capable of holding internal representation, explanations of events concerning the AI must account for this fact in some way—either to suppress this bias, or to clarify it.

The bias in attributing an internal representation to AI processes is prevalent in the general public and even domain and AI experts (Darling, 2015; Salles et al., 2020). For example, Ehsan et al. (2021) found that AI experts (computer-science students of AI curriculum) and non-experts alike, through explanations, attribute modes of human-like power of mind to AI behavior, even (though less so) when the explanations do not contain explicit information about justification behind the AI's decisions, and the effect is stronger when the explanation is given in natural language. Additionally, *concept explanations* (TCAV, Kim et al., 2018) are an explicit attribution of symbolic representation to AI (Section 4.3), and *natural-language explanations* (Narang et al., 2020; Wiegreffe & Marasovic, 2021) attempt to give AI a human voice (Section 4.4). Even the act of *text marking*, a common explanation format in XAI, can be interpreted with an anthropomorphized lens (Marzouk, 2018; Jacovi & Goldberg, 2021).[6] Finally, AI researchers and developers are susceptible to using anthromopomorphic rhetoric, as well (Watson, 2020).

On mitigating anthropomorphic bias. The attribution of human-like internal representation to AI as a result of anthropomorphic bias is implicit, possibly of subconscious

---

6. Marzouk (2018) note many possible attributions of intentionality to text marking: Marking "easy to forget" text, marking definitions, marking unclear text to investigate later, summarizing text, marking text contradictory to personal belief, and so on. When reading marked text, the perception of how this marking came to be influences how it is understood.

habit, and is therefore potentially damaging to the utility of AI explanations (Ehsan et al., 2021; Hartzog, 2015).[7] There are three possible methods of mitigating this danger: (1) To adapt to the bias by understanding the perceived power of mind, and taking action on AI design to accommodate it (Zlotowski et al., 2015); (2) to control the perception of power of mind by taking action to properly communicate the AI's capabilities (Darling, 2015); or (3) to remove it entirely by communicating to the explainee the lack power of mind in an AI (see e.g., scientific explanation of natural phenomena, such as explaining how planes fly, or explaining how tools work) (Epley et al., 2007).[8]

3.2 Categories of Folk Concepts of Behavior

Figure 2 outlines a categorization of causes that, according to empirical evidence, humans comprehend intuitively (Karniol, 1978; Knobe & Malle, 1997; Malle, 2003; Burra & Knobe, 2006): Representation causes, external causes, and internal representations, with respect to a contrast case. We describe each of these concepts and demonstrate their use in a running example.[9]

Running example (*self-driving car*, Figure 3). Consider a self-driving car that was involved in an accident: The car drove into a wall. An explanation is provided: The car had crossed the speed limit—driving at 50 km/h even though the limit was 20 km/h—due to misidentifying a nearby 20 km/h speed sign as a 50 km/h sign, because debris was covering its camera. As a result, the car had veered off-road due to an unobservable bump in the road (at which point steering became impossible), and crashed into a nearby wall.

Supposing that the explanation is "true", we assume that a human explainee considers the AI software in the car as an actor, and that they consider the explanation satisfactory. We will highlight one possible mental model that could manifest for this example.

3.2.1 Internal Representation

Internal representation refers to how the actor subjectively represents the world (independently from the actual world). E.g., "The man robbed the bank because *he needed money*".[10]

---

7. Whether humans can be "correct" in attributing mental states to AI at all is a matter of philosophical debate, but nevertheless there is sufficient evidence that humans do make this attribution often (Shelvin, 2022). We argue that explaining AI behavior successfully, so that explainees' mental models are coherent, is a goal which is independent of the discussion of whether these mental models are philosophically "correct" or not, if AI is to be useful in society.

8. In particular, human-robot interaction research discusses all three methods with respect to robots: For example, Natarajan and Gombolay (2020) conduct a user study controlling for anthropomorphic rhetoric in human-robot interaction, through personification, and feedback such as apology or indifference; finding significant effect on trust. Darling (2015) discusses anthropomorphic framing of robots, and argue for both beneficial and detrimental aspects of anthropomorphism, and aspects that control it (framing robots as tools or as companions).

9. Terms and categorizations in this section are simplified slightly from their philosophy counterparts, to reduce the barrier of entry for the AI audience.

10. There are multiple models of mental states in philosophy, the most common and simplistic being that of collections of beliefs and desires; additional models include values, emotions, thoughts, outcome-beliefs and ability-beliefs (Heider, 1958) among others (Malle, 2003; Andrews, 2006). It can be argued that promoting the attribution of beliefs and desires to automated processes encourages excessive anthropo-
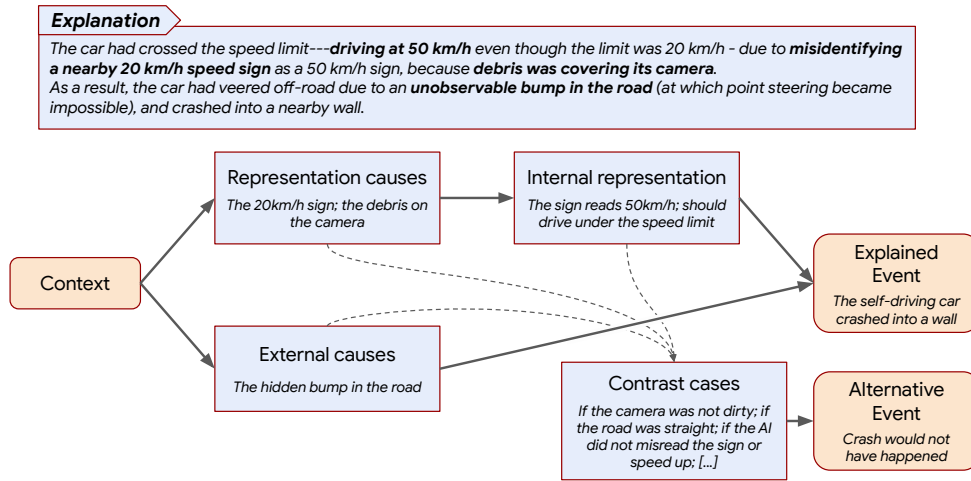
Figure 3: Schematic of the *self-driving car* example in the categorization the explanatory narrative discussed in Section 3.

Running example: The explainee may understand that the car's software internally represents the sign as a sign of a 50 *km/h* speed limit, and is representing the rule to drive at or below the speed limit.

3.2.2 Representation Causes

Representation causes refer to causes in the world that causally affected the actors' internal representation in some way (i.e., if we intervene on the representation cause, the actor's internal representation would change).[11] E.g., "The man robbed a bank because he needed money to treat *an illness*"—if the illness did not exist, the man would not need the money ("needing money" being an internal representation).

Running example: The sign of 20 *km/h*, the camera, and the debris, are all objective causes of representation, as they provide causal history to how the AI represented the sign. In other words, if one of these factors was different in some way (e.g., no debris on the camera, or sign of a roundabout ahead), our explainee would potentially expect the AI's internal representation to be different as well.

---

morphism of machines. In this work, we discuss the attribution of internal representation only to the extent of evidence that it occurs, without adopting a specific definition for what "internal representation" can refer to, as this is an active area of debate (Shelvin, 2022).

11. Representation causes and the representation itself have different roles in communicating information about the actor's behavior; for example, Brem and Rips (2000) found that evidence (objective causes) is considered more explanatory among more knowledgeable explainees in legal settings, in comparison to explicitly explaining subjective internal representation directly, than among explainees with less expertise.

3.2.3 EXTERNAL CAUSES

External causes are the objective causes in the world which are *unrelated* to an actor's internal representation. E.g., "the man successfully robbed the bank because *the security alarm was faulty*"—whether the security enabled the robbery or not, has no effect on the man's intentionality to rob the bank.

Running example: Our explainee may comprehend the unobserved bump in the road as an external cause: In the explainee's mental model of the accident, regardless of whether the road bump existed or not, the AI's *internal representation* would not change—the car's AI would still misidentify the sign, and drive at 50 km/h. However, the final event would change—the accident would not have happened—which means that the hidden bump did have a causal effect on the accident without affecting the AI's representation of the world.

3.2.4 **Contrast Cases**

Explanations, as a function of mental models, are widely accepted to be contrastive (Lugg, 1983; Lipton, 1990; Hilton, 1990). This is due to the limit of cognitive load of humans to process "complete" explanations (Lewis, 1986b; Miller, 2018), so the explanation is simplified by contrasting the event against another event of similar context (Ylikoski, 2006). As such, all of the previous categories of folk concepts (*internal representation*, *representation causes*, and *external causes*) can be comprehended with respect to the contrast case that intervenes on them.

We can make a distinction between "bifactual" and "counterfactual" contrast cases: Bifactual being an event which occurred in reality (answering "why did P happen in context A, while Q happened in context B?"), and counterfactual being a theoretical-fictional event (answering "why did P happen in context A instead of Q?") (Miller, 2018).

Running example: The debris on the camera, being a representation cause, implies a counterfactual reality: "Had the camera not been dirty, the car would have not misidentified the sign." The same information can be given via a bifactual instead: "Last week, the car had driven on the same road with a clean camera, at 20 km/h, and the accident did not occur."

3.3 Implications

The categorization of folk concepts of behavior has two relevant implications for constructing an explanatory narrative:

Representation causes and representation form a causal chain, such that explanations without *both* components are more difficult to understand. Explaining a representation cause without the resulting internal representation may force the explainee to assume what the representation is; explaining the internal representation without the causes that led to it may force the explainee to assume what those causes were. We explore this in Section 4.

The explainee may make incorrect generalizing assumptions by assuming missing components. The step of interpreting *representation causes* into *representation* by the explainee serves to apply more general rules that conform to the causal history coherently (Murphy & Medin, 1985): We attribute an internal representation to the actor based

on our knowledge of what representation *we* may have had in a similar context (Andrews, 2006). If the representation is hallucinated or misunderstood, this attribution may be wrong, and thus incoherent (Lewis, 1986a). As Nowak et al. (2013) explain, mental models of abstract, non-linear processes happening in complex systems are almost impossible to construct solely using individual cognitive capabilities.

4. The Narratives of AI Explanation Methods

Analyses of XAI methods often focus on their ability to satisfy heuristics of what explanation methods should do, and conclude that they are fragile (Kindermans et al., 2019; Hooker et al., 2019; Jacovi & Goldberg, 2020). But it remains unclear what exactly is the point of failure, in terms of the potential explainee's mental model, and the contradictions between it and observed behavior.

Using the actor-centric framework developed thus far (Section 3.2), we are now able to diagnose a given XAI method for potential contradictions between what the method communicates about model behavior, and the mental model of the explainee from which they extrapolate AI behavior.

This section is a case study of such diagnoses of four common types of AI explanation. Each diagnosis follows a general structure: (1) Description of the mechanism; (2) The possible information it communicates, in the language of folk concepts of behavior; (3) An illustrative example of the resulting perceived narrative; (4) Diagnosis of potential failure modes. We provide an overview in Table 1.

Assumptions. (A) *On internal representation:* We assume that the AIs in the case studies are perceived as actors capable of holding internal representations (see Section 3.1). (B) *On correct explanation:* We are not concerned in this section with whether the explanations are "faithfully" describing the model (see Section 2), but only in how a fictional explainee may comprehend them. (C) *On interactive explanation:* For demonstration we assume a single iteration for explanation to surface possible contradictions in the scope of the iteration. This is not to say that the explanation is "forfeit" once contradictions surface, but that *additional iterations are required* to re-establish coherence. Each iteration of explanation is a direct result of the previous iteration's mental model, which makes interactivity indispensable for its implementation. As a general approach, in all of the following cases, once an issue is found—the hypothesis could be adjusted by exploring explanations for additional examples.

4.1 Training Data Attribution

Mechanism. A class of methods for supervised AI models attempt to attribute the examples in the training data which "influenced" a particular decision. Influence functions approximate the effect of removing an example from the training data on the loss of the explained example (Koh & Liang, 2017; Han et al., 2020); Cook's distance measures the change in prediction for an example for linear regression models by removing a training example from training and re-training the model (Cook, 1977).

---

12. Gradients (Li et al., 2016; Selvaraju et al., 2020), SmoothGrad (Smilkov et al., 2017), LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), inter alia.

Table 1: Summary of Section 4 and application to various explanation methods. (∗) In standard neural networks, the contrast case is explicit for continuous-space inputs (vision, speech) but implicit for embedded inputs (discrete sequences, natural language).

Mechanism	Explicit components	Missing components	Possible failure (e.g.)
Similar training examples (§4.1)	Contrast case	Repr. cause, representation	*Contradiction with perceived repr. cause* **(A)** Explainee will look for similarities (repr. causes) between examples that are important to model behavior. *Contradiction:* The hypothesized similarities are unimportant to model behavior, s.t. model behavior and expected model behavior will be different on additional examples which share the similarities.
Influence functions (Koh & Liang, 2017) (§4.1)	Contrast case, repr. cause	Representation	*Contradiction with perceived representation* **(B)** Explainee will assume that the model learned to "represent and use" some property (repr. cause) in the influential example, where the property is a shared characteristic between the real and influential example, despite a different model representation. *Contradiction:* Model behavior will differ from expectation on additional examples with the property, due to the different internal representation.
Feature attribution[12] (§4.2)	Contrast case[(∗)], repr. cause	Contrast case[(∗)], representation	*Contradiction with perceived representation* **(B)** Explainee may assume that the model is interpreting some word in the input (representation) in a specific context (e.g., using a gender pronoun to signal gender) while the model is using it for something else (e.g., the co-referred noun of the pronoun). *Contradiction:* Model behavior will differ from expectation on examples that share the same repr. cause (the same gender pronoun), but differ in representation (the entity that the pronoun is referring to).
TCAV (Kim et al., 2018), MDL probing (Voita & Titov, 2020) (§4.3)	Representation	Contrast case, repr. cause	*Contradiction with perceived context* **(C)** Model recognizes that some property (e.g., striped fur) was in the image, but counterfactual is missing: Explainee may assume "striped fur rather than mono-color fur", but the real contrast case may be "striped fur rather than dotted fur". *Contradiction:* Model behavior will differ from expectation on examples which share properties with the hypothesized counterfactual (e.g., mono-color fur examples).
Amnesic Probing (Elazar et al., 2021a), CausalM (Feder et al., 2021) (§4.3)	Contrast case, representation	Repr. cause	*Contradiction with perceived repr. cause* **(A)** The explainee may assume that some part of the example caused the representation (e.g., whiskers in the image and the model recognizing whiskers), while the representation is based on a different repr. cause. *Contradiction:* Model behavior will differ on examples which share the real repr. cause (e.g., blades of grass), but not the perceived repr. cause (e.g., whiskers).
WT5 rationalization (§4.4)	Representation	Contrast case, repr. cause	As the function of this mechanism is the same as *concept attribution*, so are its failures.

Folk concepts. The influential examples produced by training data attribution methods can be interpreted as *representation causes*: Communicating what influences the AI's representation of the input. But interestingly, since influential examples do not communicate the contrast case, they can be perceived in two different ways: (1) As *bifactual*, i.e.,

"explanation by example." The influential example is simply another related instance of behavior; (2) or as a *counterfactual*, a contrast case in which if the influential example was not part of the AI's training, its loss function on the current example would have changed. The two different perspectives can potentially change how the explainee will understand the explanation.

Demonstration (*carnivorism prediction*). Consider the case of a classifier that classifies whether a given image of an animal is a carnivore or a non-carnivore. The AI model takes an image of a cat and outputs the decision that it is a carnivore (Figure 4a). An influence function-based method provides an explanation as a training data image of a tiger which influenced the prediction.

If interpreted as a counterfactual, an explainee may understand that the model is making a generalization for carnivores based on shared characteristics between the images (e.g., that they belong to the carnivorous felidae family; or that a striped fur is indicative of carnivorism).[13]

If interpreted as a bifactual, however, it is possible to formulate a different mental model from the *same* explanation in a different context. Suppose that the model erroneously categorizes a picture of a lion as a non-carnivore, but the explanation given is the same training-data image of a tiger (Figure 4b). The meaning of the explanation could now be perceived as an answer to the question: *Why did the model decide that the lion is not a carnivore—while the tiger is?*

Potential failure (*implicit internal representation*). The reason for the two different interpretations for the same explanation, described above, is that the explanation does not communicate what the model is representing in the relationship between the two images. Therefore, the explainee is free to make assumptions about this representation, which may be correct or incorrect. An incorrect assumption would lead to a contradiction when the model behaves differently from expected in the future.[14]

4.2 Feature Attribution

Mechanism. Feature attribution methods, whether discrete (e.g., LIME, (Ribeiro et al., 2016) and erasure (Arras et al., 2016; Feng et al., 2018)) or continuous (e.g., gradient-based (Simonyan et al., 2014) and attention flow (Abnar & Zuidema, 2020; Ethayarajh & Jurafsky, 2021)), derive which portions of the input have influence on the AI's behavior by intervening on (perturbing) the input in some systematic way and observing behavior changes.

Folk concepts. Feature attribution methods provide *representation causes*—regions in the input to the model (e.g., phrases in text or pixels in image) that especially influenced how the model represents the input. However, they traditionally do not provide information on how the models represent those regions internally.

---

13. Of course, it is also possible to interpret the explanation differently. Assume this interpretation for the sake of demonstration.

14. This issue is further exacerbated by the fact that many AI systems, in particular neural models which are explained by influence functions, do not possess a symbolic internal representation system. This complicates the exercise of making "correct" assumptions about the models' internal representations.
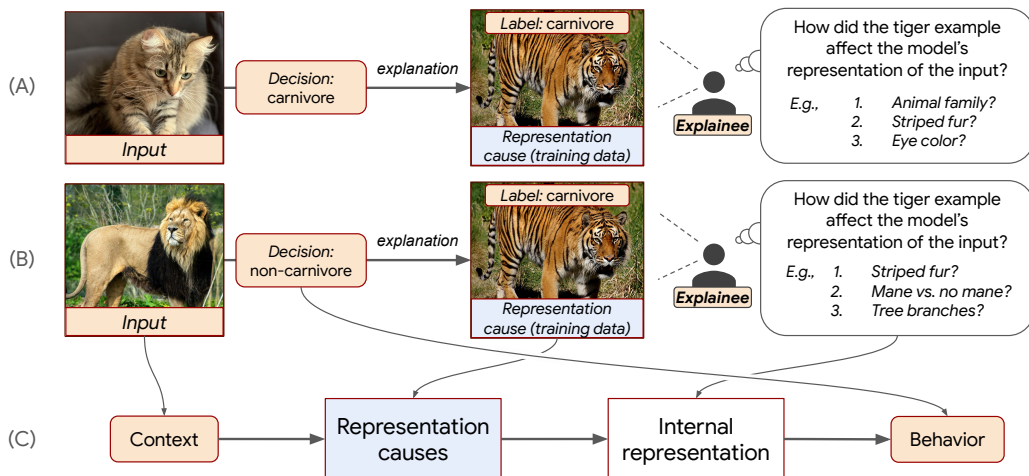
Figure 4: Demonstration of the explanatory narrative of influential examples (Section 4.1). The "influential example" explanation is the same in both contexts, yet only constitutes as a *representation cause*, and can be interpreted in very different ways by imagining the missing explanation for *how the model represents the cause*. (C) shows the incomplete mapping to the folk concepts that compose the explanatory narrative.

Demonstration (*restaurant review*). Suppose that in a *sentiment classification* task, a classifying model predicts the binary sentiment polarity of a restaurant review:

Best <u>Mexican</u> I've ever had! ⟶ *positive*

Where the underlined text is the explanation. This explanation is likely to be interpreted as a representation cause: The claim is that if this part of the input changed, then the classifier's internal representation of the input will change significantly, and therefore the decision would also change.

Potential failure (*implicit contrast case and internal representation*). The explanation is missing what the classifier is representing in the attributed phrase, and what type of intervention on this phrase would change the classifier's representation. Therefore, the explainee is at risk of assuming what these missing are. For illustration, below are two possible assumptions about this representation and contrast case:

(1) Best <u>Indian</u> I've ever had! *(country identity)*
(2) Best <u>fish</u> I've ever had! *(food category)*

This ambiguity is a potential point of fragility in the explainee's comprehension of the model's behavior. Without additional clarity on these folk concepts, the explainee may assume one of the options, and discover a contradiction if the assumption is incorrect.
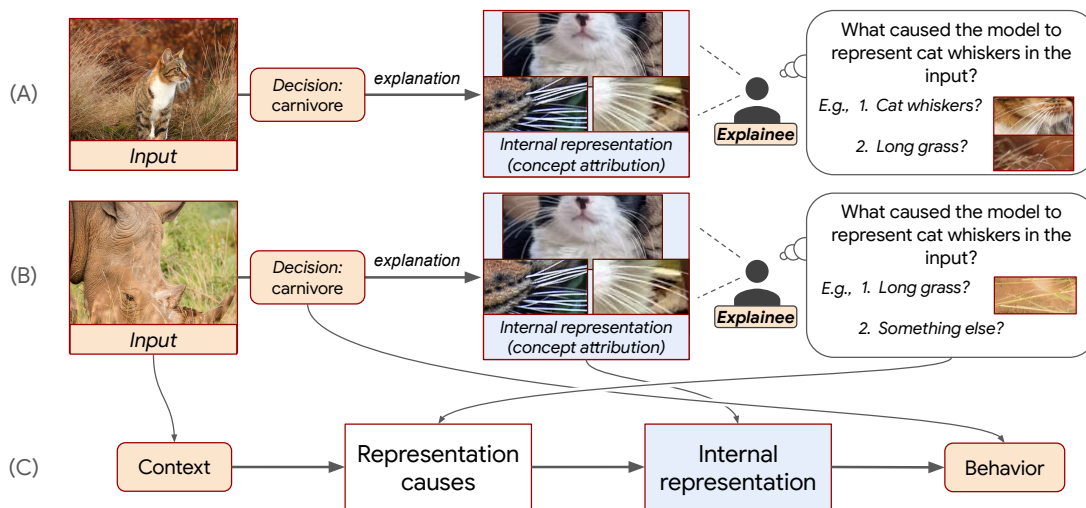
Figure 5: Example for concept explanations (Section 4.3). The explainee may hallucinate the cause of the attributed concept to be the whiskers in the image (or any particular object in the image), even though this is not part of the explanation: The explanation only communicated the internal representation of the model, but not what could have affected this representation. (C) shows the incomplete mapping to the folk concepts that compose the explanatory narrative.

4.3 Concept Attribution

Mechanism. A class of XAI methods attempt to characterize which human-interpretable abstractions (concepts) are represented by, and used in, the AI model's reasoning process. In this area, probing methods (Adi et al., 2017; Conneau et al., 2018) characterize what is encoded in the model's representation, while TCAV (Kim et al., 2018), MDL probing (Voita & Titov, 2020), amnesic probing (Elazar et al., 2021a), causal mediation analysis (Vig et al., 2020), causal abstractions (Geiger et al., 2021), inter alia, provide more insight on the role of the concepts in model behavior.

Folk concepts. Concept attribution methods map the AI model's internal representation of the input into human-interpretable concepts, therefore they attempt to communicate internal representation. Importantly, whether a concept is detected in the internal representation does not entail whether the concept really does or does not exist in the context.

Demonstration (*whiskers attribution*). Suppose that a model decision, classifying a cat image as a carnivore, has been attributed with the "cat whiskers" concept (Figure 5a). The concept is commonly defined as a set of samples that contain the concept.

Potential failure 1 (*implicit representation causes*). Current methods of concept explanations communicate internal representation without its causes. This means, for example, that the explainee may understand that the model represents that the image of a cat has whiskers, but not necessarily what caused this.

This is fragile point in the explanatory narrative: The explainee may make an assumption about what caused the representation of the concept, and this assumption may not be true. For example, if the image indeed has a cat with whiskers, the explainee may assume that the model's representation of the whiskers concept is caused by the whiskers in the image—when in reality, perhaps the model mistook blades of long grass in the background of the image for whiskers. The assumption can cause a failure of coherence if the model behaved similarly on other images which do not have whiskers, but do have similar blades of grass (Figure 5b).

Potential failure 2 (*implicit contrast case*). In the case of classic *probing* methods which communicate whether a concept is being represented by the model, it is possible that this representation is not a cause of the model's final decision (i.e., it does not explain the decision). This is because the counterfactual case, where the concept is absent, is not part of the explanation.

This has been a subject of recent criticism for probing methods, on the basis of "correlation does not equal causation"—where although probing methods infer that the model represents some concept, no guarantee is given if the model actually uses this concept to make its decisions (Tamkin et al., 2020; Geiger et al., 2020; Ravichander et al., 2021). This has led to the development of a causally-informed class of methods (Vig et al., 2020; Feder et al., 2021; Geiger et al., 2021) which provide a stronger guarantee that causality is correctly attributed. This can be accomplished, for example, by showing that the model changes its decision if it ceases to recognize the concept via a counterfactual (Elazar et al., 2021a; Feder et al., 2021).

4.4 Natural-language Generation (a.k.a. Abstractive Rationales)

Mechanism. Models generating "rationalizations" as natural-language explanations (Ehsan et al., 2018; Wiegreffe et al., 2020; Narang et al., 2020) learn from human-written explanations to produce a natural text from the AI model's hidden representation, attempting to justify their actions, inspired by the way that a human would explain their own behavior (Wiegreffe & Marasovic, 2021).

Folk concepts. This class of explanations attempts to communicate what the model is representing in natural language, therefore they communicate the model's *internal representation*. Note that this is a very similar narrative function to *concept attribution* (Section 4.3). The medium of natural-language communication may reinforce anthropomorphic bias in comparison to other mediums (Ehsan et al., 2021).

Demonstration. Continuing the *whiskers attribution* example from Figure 5, such a model may generate the explanation: "*Because it has whiskers*", "*because it has stripes*" or even "*because it eats meat*" as a rationalization.

Potential failures. Natural language rationalization communicates the same folk concepts of behavior as concept attribution, therefore it shares the same potential coherence failures (for example, implicit *representation causes*), despite these two methodologies having very different underlying technologies.

Table 2: Various, seemingly different, XAI methods may share the same failures according to our abstraction (Section 5).

Incompleteness type		Potential failure	Mechanisms
(A)	Missing representation cause	Contradiction with perceived representation causes	Bifactual training examples, concept attribution
(B)	Missing internal representation	Contradiction with perceived representation	Training data attribution, feature attribution
(C)	Missing contrast case	Contradiction with perceived contrast case	Feature attribution, concept attribution

5. Towards Successful Explanations

We summarize the main implications from the analysis thus far:

Explanations should establish an explanatory narrative which explicitly communicates all relevant folk concepts of behavior. The underlying root issue in all potential failures discussed in Section 4 is *an under-specification of the AI process by the explanation* (Table 2). Unaccounted components of the explanatory narrative are at risk of being "filled in" by the explainee through potentially incorrect assumptions, leading to contradictions. The explanatory narrative we propose is as follows: "Something" in the context (input data, training data, or algorithm; *representation causes*) caused the AI to represent "something" (*internal representation*) which affected the explained outcome; intervening on the representation causes will change the representation, ultimately changing the outcome (*contrast case*). Additional relevant causes which had no effect on the AI's representation, but nevertheless affected the outcome (*external causes*) should be explicitly marked as such, with additional contrast cases. See an illustrative example in Figure 6.

Explanations should use interactivity to resolve contradictions. In this work we regard something as "successfully" explained if the explainee can establish a coherent mental model, without observable contradictions to it. But explanations do not necessarily need to accomplish this in "one shot", as humans naturally use interactivity to adjust incoherent mental models. Therefore, we stand to make breakthroughs in successfully explaining AI not only by improving the explanatory narrative that the explanation communicate, but also by allowing the explainee to test their hypothesis via interactive interrogation of the AI (Gehrmann et al., 2019; Gehrmann, 2020; Krarup et al., 2021).

Additional research is required on *explainee profiling* (Fischer, 2000; Johnson & Taatgen, 2005) to characterize how different explainees may construct mental models differently. The definition of a successful explanation, as a function of a coherent mental model, is a definition that involves the explainee. In order to understand the mental model of the explainee, we must establish who the explainee is, and what prior knowledge they may leverage in their assumptions. Currently, explainee profiling in XAI is often limited to familiarity with AI technology or expertise at the end-task (AI experts/novices", "domain experts", "data scientists"; e.g.,Strobelt et al., 2017; Hohman et al., 2019; Kaur
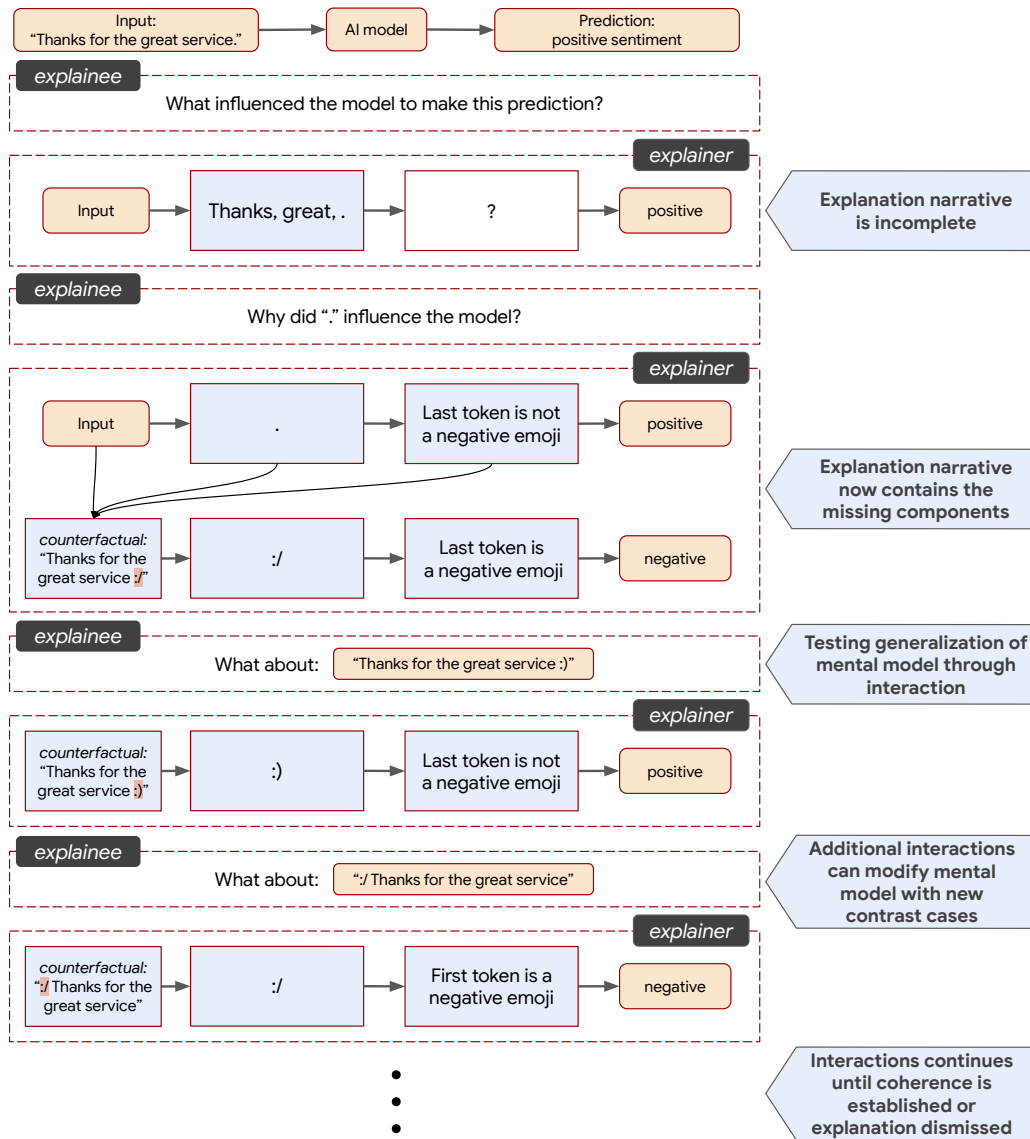
Figure 6: Illustrative example of how interactive interrogation and a complete explanatory narrative can serve as modes of explanation (Section 5).

et al., 2020; Ehsan et al., 2021), but additional research may uncover other important properties of user models, such as cognitive or social properties.

6. Discussion

On scientific explanation and argumentation. As mentioned, this work is concerned with communicating information about the process that led to AI behavior, rather than justifying this behavior or verifying its correctness (argumentation) (Kakas & Michael,

2020; Fok & Weld, 2023; Miller, 2023; Langley, 2019). The process of explaining behavior is partially discussed by literature on scientific explanation (Woodward & Ross, 2021), though this literature is generally more concerned with the specifics of deriving causal insights about processes, while this work is concerned with how (or what it would require) to *properly communicate* these derived insights to humans, specifically in the case of AI behavior.

On external causes. External causes are one of the central building blocks of an explanatory narrative explored in this work. However, the explanation methods and failure cases discussed in Section 4 do not mention external causes. Why is this the case? The AI and XAI utilities discussed here pertain to the AI mechanism itself, outside of a context in which external factors can have an effect. For example, in a robotics use-case, external hardware factors (such as water damage or wear) can affect behavior, and thus have a place in an explanatory narrative. However, such factors are detached from XAI methods that only attempt to explain the AI component.

7. Conclusion

This work identifies two different perspectives of explanation: (1) What the explanation method is communicating about the AI behavior; (2) what the explainee actually comprehends about AI behavior from the explanation. We find that the explainee may derive incorrect generalizing rules about AI behavior, causing a mismatch between (1) and (2), if the explanation is unintuitive or insufficient.

Erroneous generalizing assumptions will cause contradictions to manifest between *additional* AI behavior and the explainee's mental model. In the event of observed contradictions, we say that the mental model is incoherent, and that coherency is a primary attribute of good explanation. Successfully explaining without contradictions does not necessarily require a "perfect" initial explanation, since contradictions can be resolved via interactive interrogation of AI behavior, iteratively adjusting the mental model until it is coherent.

We apply this framework to a variety of XAI methods, and find that contradictions systematically arise from missing information in the explanation (in terms of how humans comprehend explanations: Through representation causes, internal representation, external causes and a contrast case). This provides us with a path forward towards the design of XAI methods that can be said to provide coherent explanation, specifically by being complete and interactive.

Extensions and future work. We note additional exceptionally multi-disciplinary research directions related to successful explanation: 1. How to communicate lack of power of mind in what society considers as AI, or "intelligent" automated processes. As noted in Section 3.1, this question is discussed in *human-robot interaction*, but remains an open question in other settings. 2. Characterizing the budget sufficient to proving that an explanation is coherent—subject to a particular use-case. 3. The research and integration of additional social science sources on theory of XAI communication with humans: Discourse theory (Macdonell, 1986); collaboration theory (Salas et al., 2017); and other cognitive habits in comprehending explanations of behavior, e.g., the least effort principle (Zipf, 1950), confirmation bias (Nickerson, 1998), and belief bias (Gonzalez et al., 2021).

Acknowledgments

Appendix A. Current XAI Desiderata as Measures of Coherence

As mentioned in Section 2, human society uses coherence as a measure to rely on explanations, due to intractability in proving correctness. Interestingly—though perhaps unsurprisingly—this narrative can also be applied to the development of the relaxed measures of faithfulness in the XAI community. In this appendix section we show that current XAI measures of faithful explanations can be positioned, with some reservation, as measures of coherence.

Neighborhood similarity (e.g., Alvarez-Melis & Jaakkola, 2018; Yin et al., 2021; Ding & Koehn, 2021) measures the degree to which similar events are explained similarly. Failure here (i.e., dissimilarity) can be interpreted as a contradiction, under the assumption that the explanation should generalize to examples in the neighborhood.

This is a relaxed measure of coherence which only tests for contradictions in a neighborhood of contexts, and assumes that the explanation is a proxy for the explainee's mental model.

Model similarity (e.g., Wiegreffe & Pinter, 2019; Ding & Koehn, 2021) measures the degree to which two models with similar *behavior* are explained similarly. One can also define measures based on model dissimilarity for models which behave very differently (Adebayo et al., 2018).

This measure is a variant of the *neighborhood similarity* above, which expands the contradiction search space, and assumes that the two models' explanations will communicate the same mental model to the explainee.

Fidelity (Ribeiro et al., 2016; Guidotti et al., 2018) measures the degree to which a simpler, "explainable" surrogate model is able to mimic the black-box model. In this case, the explanation of the black-box model is the simpler model.[15]

This measure is a direct adaptation of coherence: The simple model serves as the hypothesis. The budget for proving or refuting coherence can be formalized as the breadth and depth of search for possible instances for which the surrogate model fails to mimic the explained model. However, the required level of fidelity (i.e., quantity of contradictions) is challenging to relate to theory of mind literature. Empirical XAI studies that aim to connect user trust to explanation fidelity found that the way explanations are presented and the underlying model accuracy often overshadow the effect of fidelity, thus making it hard to draw conclusions from the perspective of explainees (Papenmeier et al., 2019; Larasati et al., 2020).

---

15. *Fidelity* can also be considered a special case of model similarity.

Additionally, some methods of "surrogate model" explanations that report fidelity only attempt to mimic the black-box model locally around a particular instance of behavior. Such methods have a weaker connection to coherence, since they do not attempt to fit model behavior across the possible input space.

Relaxed ground truth evaluation (e.g., Sippy et al., 2020; Zhang et al., 2021a; Carmichael & Scheirer, 2021; Bastings et al., 2021; Zhou et al., 2021) defines a ground truth on "correct" explanation by explaining processes which are guaranteed, or are very likely, to reason in a particular way (e.g., a biased model designed to err systematically, or introducing a "watermark" to the data which is perfectly correlated with a label; see Zhou et al. (2021), Bastings et al. (2021)).

The connection to coherence is straightforward—the explanations are measured via the degree of accuracy to the ground truth—but notably, the empirical budget for proof of coherence manifests in the observed space of AI behavior for which the ground truth exists. For example, evaluating via watermarking only carries real weight for the space of examples with the watermark.

Simulatability (e.g., Doshi-Velez & Kim, 2017; Hase et al., 2020; Hase & Bansal, 2020) measures the ability of human explainees to simulate the AI process in a particular setting.

Simulatability is a sub-case of coherence: Where coherence measures the presence of contradictions to the mental model in all abstract meanings of this definition, simulatability tests for contradictions strictly at the final decision level. Therefore a failure by the user to predict the AI is a clear sign that a contradiction *exists* between the user's mental model and the AI, but it may not be clear what the contradiction is through simulatability alone.

Appendix B. Criticism: On Decision-level (local) and Model-level (global) Explanations

XAI literature commonly categorizes explanations into two groups: Explaining singular decisions (decision explanations, local explanations) and explaining the entire scope of model behavior (model explanations, global explanations) (Belinkov & Glass, 2019; Burkart & Huber, 2021; Setzu et al., 2021). This gives a taxonomy of explanation mechanisms, unrelated to the mental model of a particular explainee.

In this appendix, we scrutinize the utility of this categorization: Is the categorization of decision and model explanations potentially descriptive of any differences in the explainee's mental model?

Decision-level explanations and coherence. Decision explanations, in themselves, by definition are not constrained with coherence, since they only explain individual instances of behavior. However, this does not mean that they are not *perceived to be* describing generalizing behavior.

Indeed, under the framework of coherence, explanation is inherently an attempt to communicate generalizing rules. Decision level explanations should be considered as modes of communicating information which can apply beyond the explained instance of behavior.

Given this conclusion, we argue that "decision-level" categorization is potentially *misleading* as a description of explanation methods. This argument has also been discussed by Hoffman et al. (2020).

Is the decision-level and model-level categorization descriptive of the function of XAI methods? Both decision-level and model-level explanations can communicate information about representation causes, internal representation, external causes, as well as counterfactual and bifactual information directly. However, they aim to explain different events: In decision explanations, the event is the final decision of the AI on a particular instance. But model-level explanations can potentially explain two different events:

1. The event can be the model itself as the outcome of the process that created it. For example, characterizing the functionality of different components in a compositional neural network (Subramanian et al., 2020) or the different kernels in a convolutional neural network (Zeiler & Fergus, 2014) explains the model by building a counterfactual context which would have resulted in a different model.

2. The event can be the aggregation of the model's behavior on a large collection of instances, making it an aggregating case of decision-level explanations. For example, in explaining that a model achieves strong performance on some task because of exploiting a spurious heuristic (Gururangan et al., 2018), the "contrast case" is a reality where the model is the same, but the *instance space* is different (from instances that exhibit the heuristic, to instances that do not)—such that its *decisions* would be different in this instance space, compared to the previous decisions (e.g., Elazar et al., 2021b; Rosenman et al., 2020; McCoy et al., 2019).

The two different types of events carry different implications on what the explainee may understand about the AI. For example, the contrast case between the two events is different: In (1) it is a different model, while in (2), it is the same model deployed in different contexts.

And yet, the same denomination of "model-level explanations" refers to both perspectives interchangeably in the literature (e.g., Zhang et al., 2021b). Therefore it can be interpreted as an ambiguous or confusing term, and not descriptive of how the explainee will interpret a given explanation.

References

Abnar, S., & Zuidema, W. H. (2020). Quantifying attention flow in transformers. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. R. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4190–4197. Association for Computational Linguistics.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *CoRR, abs/1810.03292*.

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *CoRR, abs/1806.08049*.

Amir, O., Doshi-Velez, F., & Sarne, D. (2019). Summarizing agent strategies. *Auton. Agents Multi Agent Syst., 33*(5), 628–644.

Andrews, K. (2006). The functions of folk psychology..

Arras, L., Horn, F., Montavon, G., Müller, K., & Samek, W. (2016). "what is relevant in a text document?": An interpretable machine learning approach. *CoRR*, *abs/1612.07843*.

Baan, J., ter Hoeve, M., van der Wees, M., Schuth, A., & de Rijke, M. (2019). Do transformer attention heads provide transparency in abstractive summarization?. *CoRR*, *abs/1907.00570*.

Bastings, J., Ebert, S., Zablotskaia, P., Sandholm, A., & Filippova, K. (2021). "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification..

Belinkov, Y., & Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72.

Bradley, D. (2017). Should explanations omit the details?. *British Journal for the Philosophy of Science*, *71*.

Brem, S., & Rips, L. (2000). Explanation and evidence in informal argument. *Cognitive Science*, *24*, 573–604.

Bucinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In Paternò, F., Oliver, N., Conati, C., Spano, L. D., & Tintarev, N. (Eds.), *IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020*, pp. 454–464. ACM.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.*, *70*, 245–317.

Burra, A., & Knobe, J. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture*, *6*(1-2), 113 – 132.

Carmichael, Z., & Scheirer, W. J. (2021). On the objective evaluation of post hoc explainers. *CoRR*, *abs/2106.08376*.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In Gurevych, I., & Miyao, Y. (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2126–2136. Association for Computational Linguistics.

Cook, R. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in hci agents. *Computers in Human Behavior*, *29*(3), 577–579.

Dacey, M. (2017). Anthropomorphism as cognitive bias. *Philosophy of Science*, *84*(5), 1152–1164.

Darling, K. (2015). 'who's johnny?' anthropomorphic framing in human-robot interaction, integration, and policy..

Ding, S., & Koehn, P. (2021). Evaluating saliency methods for neural language models. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., & Zhou, Y. (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 5034–5052. Association for Computational Linguistics.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning..

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2018). Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, p. 81–87, New York, NY, USA. Association for Computing Machinery.

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M. J., & Riedl, M. O. (2021). The who in explainable AI: how AI background shapes perceptions of AI explanations. *CoRR, abs/2107.13509*.

Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021a). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics, 9*, 160–175.

Elazar, Y., Zhang, H., Goldberg, Y., & Roth, D. (2021b). Back to square one: Bias detection, training and commonsense disentanglement in the winograd schema. *CoRR, abs/2104.08161*.

Epley, N., Waytz, A., & Cacioppo, J. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886.

Ethayarajh, K., & Jurafsky, D. (2021). Attention flows are shapley value explanations. In Zong, C., Xia, F., Li, W., & Navigli, R. (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pp. 49–54. Association for Computational Linguistics.

Feder, A., Oved, N., Shalit, U., & Reichart, R. (2021). Causalm: Causal model explanation through counterfactual language models. *Comput. Linguistics, 47*(2), 333–386.

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., & Boyd-Graber, J. (2018). Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Fischer, G. (2000). User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction, 11*.

Fok, R., & Weld, D. S. (2023). In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making..

Gehrmann, S. (2020). *Human-AI Collaboration for Natural Language Generation with Interpretable Neural Networks*. Ph.D. thesis, Harvard University.

Gehrmann, S., Strobelt, H., Krüger, R., Pfister, H., & Rush, A. M. (2019). Visual interaction with deep learning models through collaborative semantic inference. *CoRR*, *abs/1907.10739*.

Gehrmann, S., Strobelt, H., Krüger, R., Pfister, H., & Rush, A. M. (2020). Visual interaction with deep learning models through collaborative semantic inference. *IEEE Trans. Vis. Comput. Graph.*, *26*(1), 884–894.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *CoRR*, *abs/2106.02997*.

Geiger, A., Richardson, K., & Potts, C. (2020). Neural natural language inference models partially embed theories of lexical entailment and negation. In Alishahi, A., Belinkov, Y., Chrupala, G., Hupkes, D., Pinter, Y., & Sajjad, H. (Eds.), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pp. 163–173. Association for Computational Linguistics.

Ghorbani, A., Abid, A., & Zou, J. Y. (2019). Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 3681–3688. AAAI Press.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. A., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In Bonchi, F., Provost, F. J., Eliassi-Rad, T., Wang, W., Cattuto, C., & Ghani, R. (Eds.), *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pp. 80–89. IEEE.

Gonzalez, A. V., Rogers, A., & Søgaard, A. (2021). On the interaction of belief bias and explanations. In Zong, C., Xia, F., Li, W., & Navigli, R. (Eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, Vol. ACL/IJCNLP 2021 of *Findings of ACL*, pp. 2930–2942. Association for Computational Linguistics.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, *51*(5).

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. In Walker, M. A., Ji, H., & Stent, A. (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics.

Han, X., Wallace, B. C., & Tsvetkov, Y. (2020). Explaining black box predictions and unveiling data artifacts through influence functions. In Jurafsky, D., Chai, J., Schluter, N.,

& Tetreault, J. R. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 5553–5563. Association for Computational Linguistics.

Hanisch, K. A., Kramer, A. F., & Hulin, C. L. (1991). Cognitive representations, control, and understanding of complex systems: a field study focusing on components of users' mental models and expert/novice differences. *Ergonomics*, *34*(8), 1129–1145.

Hartzog, W. (2015). Unfair and deceptive robots. *Maryland Law Review*, *74*, 785.

Hase, P., & Bansal, M. (2020). Evaluating explainable AI: which algorithmic explanations help users predict model behavior?. *CoRR*, *abs/2005.01831*.

Hase, P., Zhang, S., Xie, H., & Bansal, M. (2020). Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language?. In Cohn, T., He, Y., & Liu, Y. (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Vol. EMNLP 2020 of *Findings of ACL*, pp. 4351–4367. Association for Computational Linguistics.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259.

Heider, F. (1958). The psychology of interpersonal relations..

Hilton, D. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*, 65–81.

Hoffman, R. R., Clancey, W. J., & Mueller, S. T. (2020). Explaining AI as an exploratory process: The peircean abduction model. *CoRR*, *abs/2009.14795*.

Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2019). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Trans. Vis. Comput. Graph.*, *25*(8), 2674–2693.

Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

Islam, S. R., Eberle, W., Ghafoor, S. K., & Ahmed, M. (2021). Explainable artificial intelligence approaches: A survey. *CoRR*, *abs/2101.09429*.

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. R. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4198–4205. Association for Computational Linguistics.

Jacovi, A., & Goldberg, Y. (2021). Aligning faithful interpretations with their social attribution. *Trans. Assoc. Comput. Linguistics*, *9*, 294–310.

Johnson, A., & Taatgen, N. (2005). *User modeling.*, pp. 424–438.

Johnson, D. K. (2018). *Anthropomorphic Bias*, chap. 69, pp. 305–307. John Wiley & Sons, Ltd.

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference.. *Psychological review*, *109*(4), 646.

Kakas, A. C., & Michael, L. (2020). Abduction and argumentation for explainable machine learning: A position survey. *CoRR*, *abs/2010.12896*.

Karniol, R. (1978). Children's use of intention cues in evaluating behavior.. *Psychological Bulletin*, *85(1)*, 76–85.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–14, New York, NY, USA. Association for Computing Machinery.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Dy, J. G., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 2673–2682. PMLR.

Kim, M.-Y., Atakishiyev, S., Babiker, H. K. B., Farruque, N., Goebel, R., Zaïane, O. R., Motallebi, M.-H., Rabelo, J., Syed, T., Yao, H., & Chun, P. (2021). A multi-component framework for the analysis and design of explainable artificial intelligence. *Machine Learning and Knowledge Extraction*, *3*(4), 900–921.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). *The (Un)reliability of Saliency Methods*, pp. 267–280. Springer International Publishing, Cham.

Kirchler, M., Graf, M., Kloft, M., & Lippert, C. (2021). Explainability requires interactivity. *CoRR*, *abs/2109.07869*.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4), 507–531.

Knobe, J., & Malle, B. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In Precup, D., & Teh, Y. W. (Eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for pytorch..

Krarup, B., Krivic, S., Magazzeni, D., Long, D., Cashmore, M., & Smith, D. E. (2021). Contrastive explanations of plans through model restrictions. *J. Artif. Intell. Res.*, *72*, 533–612.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In Conitzer, V., Hadfield, G. K., & Vallor, S.

(Eds.), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pp. 131–138. ACM.

Langley, P. (2019). Varieties of explainable agency. In *ICAPS Workshop on Explainable AI Planning*.

Larasati, R., Liddo, A. D., & Motta, E. (2020). The effect of explanation styles on user's trust. In *IUI 2020 Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies*.

Lehrer, K. (1990). *Theory of Knowledge*. Westview Press.

Lei, T., Barzilay, R., & Jaakkola, T. S. (2016). Rationalizing neural predictions. In Su, J., Carreras, X., & Duh, K. (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 107–117. The Association for Computational Linguistics.

Lewis, C. (1986a). A model of mental model construction. In Mantei, M. M., & Orbeton, P. (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 1886, Boston, Massachusetts, USA, April 13-17, 1986*, pp. 306–313. ACM.

Lewis, D. K. (1986b). Causal explanation. In Lewis, D. (Ed.), *Philosophical Papers Vol. Ii*, pp. 214–240. Oxford University Press.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 681–691, San Diego, California. Association for Computational Linguistics.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, *27*, 247–266.

Lipton, Z. C. (2018). The mythos of model interpretability. *Commun. ACM*, *61*(10), 36–43.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470.

Lugg, A. (1983). Alan garfinkel forms of explanation: Rethinking the questions in social theory (new haven, ct: Yale university press 1981). pp. 186. $16.00.. *Canadian Journal of Philosophy*, *13*(4), 633–646.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4765–4774.

Macdonell, D. (1986). *Theories of discourse : an introduction / Diane Macdonell*. B. Blackwell Oxford ; New York, NY.

Malle, B. F. (2003). Folk theory of mind: Conceptual foundations of social cognition..

Marzouk, Z. (2018). Text marking: A metacognitive perspective..

Mayes, G. R. (2022). Theories of explanation..

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Korhonen, A., Traum, D. R., & Màrquez, L. (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 3428–3448. Association for Computational Linguistics.

Miller, T. (2018). Contrastive explanation: A structural-model approach. *CoRR*, *abs/1811.03163*.

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

Miller, T. (2023). Explainable ai is dead, long live explainable ai! hypothesis-driven decision support..

Morton, A. (1980). *Frames of Mind: Constraints on the Common-Sense Conception of the Mental.* Oxford University Press.

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Narang, S., Raffel, C., Lee, K., Roberts, A., Fiedel, N., & Malkan, K. (2020). Wt5?! training text-to-text models to explain their predictions..

Natarajan, M., & Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, p. 33–42, New York, NY, USA. Association for Computing Machinery.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nowak, A., Rychwalska, A., & Borkowski, W. (2013). Why simulate? to develop a mental model. *J. Artif. Soc. Soc. Simul.*, *16*(3).

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *IJCAI 2019 Workshop on Explainable Artificial Intelligence (xAI)*, *abs/1907.12652*.

Payne, S. (2003). *Users' Mental Models: The Very Ideas*, pp. 135–156.

Ravichander, A., Belinkov, Y., & Hovy, E. H. (2021). Probing the probing paradigm: Does probing accuracy entail task relevance?. In Merlo, P., Tiedemann, J., & Tsarfaty, R. (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 3363–3377. Association for Computational Linguistics.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, *abs/1602.04938*.

Rosenman, S., Jacovi, A., & Goldberg, Y. (2020). Exposing shallow heuristics of relation extraction models with challenge data. In Webber, B., Cohn, T., He, Y., & Liu, Y.

(Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 3702–3710. Association for Computational Linguistics.

Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *CoRR, abs/1811.10154*.

Rutjes, H., Willemsen, M., & IJsselsteijn, W. (2019). Considerations on explainable ai and users' mental models. In *Where is the Human? Bridging the Gap Between AI and HCI*, United States. Association for Computing Machinery, Inc. CHI 2019 Workshop : Where is the Human? Bridging the Gap Between AI and HCI ; Conference date: 04-05-2019 Through 04-05-2019.

Salas, E., Rico, R., & Passmore, J. (2017). *The Psychology of Teamwork and Collaborative Processes*, chap. 1, pp. 1–11. John Wiley & Sons, Ltd.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in ai. *AJOB Neuroscience, 11*, 88 – 95.

Sellars, W. (1963). *Science, Perception and Reality*. New York: Humanities Press.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis., 128*(2), 336–359.

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). Glocalx - from local to global explanations of black box AI models. *Artif. Intell., 294*, 103457.

Shelvin, H. (2022). Uncanny believers: Uncanny believers: chatbots, beliefs, and folk psychology...

Shvo, M., Klassen, T. Q., & McIlraith, S. A. (2022). Resolving misconceptions about the plans of agents via theory of mind. In *International Conference on Automated Planning and Scheduling*.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Bengio, Y., & LeCun, Y. (Eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Sippy, J., Bansal, G., & Weld, D. S. (2020). Data staining: A method for comparing faithfulness of explainers..

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *CoRR, abs/1706.03825*.

Sokol, K., & Flach, P. A. (2020). One explanation does not fit all. *Künstliche Intell., 34*(2), 235–250.

Sreedharan, S., Chakraborti, T., & Kambhampati, S. (2021). Foundations of explanations as model reconciliation. *Artificial Intelligence, 301*, 103558.

Strobelt, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H., & Rush, A. M. (2018). Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *CoRR, abs/1804.09299*.

Strobelt, H., Gehrmann, S., Pfister, H., & Rush, A. M. (2017). Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, *24*(1), 667–676.

Subramanian, S., Bogin, B., Gupta, N., Wolfson, T., Singh, S., Berant, J., & Gardner, M. (2020). Obtaining faithful interpretations from compositional neural networks. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. R. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 5594–5608. Association for Computational Linguistics.

Tamkin, A., Singh, T., Giovanardi, D., & Goodman, N. D. (2020). Investigating transferability in pretrained language models. In Cohn, T., He, Y., & Liu, Y. (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Vol. EMNLP 2020 of *Findings of ACL*, pp. 1393–1401. Association for Computational Linguistics.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., & Yuan, A. (2020). The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models..

Thagard, P. (1988). *Computational Philosophy of Science*. The MIT Press.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. M. (2020). Causal mediation analysis for interpreting neural NLP: the case of gender bias. *CoRR*, *abs/2004.12265*.

Voita, E., & Titov, I. (2020). Information-theoretic probing with minimum description length. In Webber, B., Cohn, T., He, Y., & Liu, Y. (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 183–196. Association for Computational Linguistics.

Watson, D. (2020). *The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence*, pp. 45–65.

Wiegreffe, S., & Marasovic, A. (2021). Teach me to explain: A review of datasets for explainable NLP. *CoRR*, *abs/2102.12060*.

Wiegreffe, S., Marasovic, A., & Smith, N. A. (2020). Measuring association between labels and free-text rationales. *CoRR*, *abs/2010.12762*.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China. Association for Computational Linguistics.

Williams, J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive science*, *34*, 776–806.

Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part i: A counterfactual account. *Noûs*, *37*(1), 1–24.

Woodward, J., & Ross, L. (2021). Scientific Explanation. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 edition). Metaphysics Research Lab, Stanford University.

Yin, F., Shi, Z., Hsieh, C., & Chang, K. (2021). On the faithfulness measurements for model interpretations. *CoRR*, *abs/2104.08782*.

Ylikoski, P. (2006). *The Idea of Contrastive Explanandum*, pp. 27–42.

Yu, M., Chang, S., Zhang, Y., & Jaakkola, T. S. (2019). Rethinking cooperative rationalization: Introspective extraction and complement control. In Inui, K., Jiang, J., Ng, V., & Wan, X. (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 4092–4101. Association for Computational Linguistics.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In Fleet, D. J., Pajdla, T., Schiele, B., & Tuytelaars, T. (Eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, Vol. 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer.

Zhang, W., Huang, Z., Zhu, Y., Ye, G., Cui, X., & Zhang, F. (2021a). On sample based explanation methods for NLP: faithfulness, efficiency and semantic evaluation. In Zong, C., Xia, F., Li, W., & Navigli, R. (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 5399–5411. Association for Computational Linguistics.

Zhang, Y., Tiño, P., Leonardis, A., & Tang, K. (2021b). A survey on neural network interpretability. *IEEE Trans. Emerg. Top. Comput. Intell.*, *5*(5), 726–742.

Zhou, Y., Booth, S., Ribeiro, M. T., & Shah, J. (2021). Do feature attribution methods correctly attribute features?. *CoRR*, *abs/2104.14403*.

Zipf, G. K. (1950). Human behavior and the principle of least effort. cambridge, (mass.): Addison-wesley, 1949, pp. 573. *Journal of Clinical Psychology*, *6*(3), 306–306.

Zlotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics*, *7*, 347–360.