Explanation in artificial intelligence: Insights from the social sciences

Ninad Hagi

April 2025

Abstract

Explainable Artificial Intelligence (XAI) has gained renewed interest as increasingly complex algorithms enter high-stakes domains such as healthcare, finance, and autonomous vehicles. This thesis examines Tim Miller's argument that insights from social science, specifically from philosophy, psychology, and cognitive studies can guide the design of more human-centered AI explanations. Conventional approaches in XAI often focus on technical or statistical transparency, yet Miller's work demonstrates that individuals look for explanations framed around meaningful causes and contrasting alternatives. His analysis highlights how people naturally interpret unexpected events by focusing on abnormal conditions, using conversational shortcuts, and forming explanations grounded in beliefs, desires, and intentions. Applying these findings, this thesis discusses why a purely probabilistic rationale rarely satisfies users' explanatory needs and how AI systems could instead offer contrastive answers and context-aware responses. Through a critical review, the thesis positions Miller's perspective as a blueprint for integrating causal structures, social cues, and iterative dialogue into AI explanation mechanisms. By synthesizing these interdisciplinary insights, the work aspires to bridge the gap between algorithmic complexity and a user's quest for understandable, trustworthy models. This human-centric perspective on AI, grounded in social science, ultimately supports more intuitive, effective, and socially acceptable interactions with intelligent systems.

1 Introduction and Overview

Introduction

In recent years, the field of artificial intelligence (AI) has witnessed a surge of interest in explainability, largely because opaque or "black-box" models can erode users' trust and hinder the effective adoption of AI systems. This renewed focus on "explainable AI" (XAI) has prompted researchers to consider not only how intelligent agents make decisions but also how those decisions can be communicated transparently to human stakeholders. Tim Miller's article, "Explanation

in Artificial Intelligence: Insights from the Social Sciences," enters this discussion by arguing that approaches to XAI should be grounded in how humans naturally craft and understand explanations. Drawing on psychological, cognitive, and social theories of explanation, Miller proposes that AI researchers can benefit from well-established findings about human reasoning, particularly the ways people perceive causes, contextualize events, and convey explanations in everyday settings.

At the highest level, the article examines why explanations matter for AI in the first place (e.g., fostering user trust, facilitating learning, or supporting ethical accountability), then reviews key insights from philosophy and social science on what makes an explanation useful. In doing so, Miller underscores that effective AI explanations go beyond mere statistical accounts of how a model arrived at its output. Instead, systems should offer explanations that align with users' cognitive preferences, such as emphasizing contrastive details, identifying abnormal factors, and tailoring the explanation to the knowledge and goals of the intended audience. Thus, Miller's core research questions focus on how time-tested ideas about human explanation, things like causal attribution, social context and conversational cues, can be integrated into AI design. By reviewing and synthesizing these theoretical perspectives, he highlights that transparency is not just a technical puzzle but a human-centered challenge requiring interdisciplinary approaches. Through this lens, the paper expands our understanding of what "explainability" entails in AI and outlines practical avenues for bridging theory and implementation.

Overview

Miller's chief argument is that merely offering more data or increased transparency at a technical level (for instance, disclosing model weights or neural network activations) is often insufficient when the goal is to satisfy a human audience.[1] [1] While it is certainly helpful for developers or AI experts to have deeper access to a system's inner workings, the broader group of non-specialist users, such as medical professionals, managers, or everyday consumers, needs something different: explanations that mesh with everyday modes of understanding. In other words, AI systems should not just reveal probabilities or numeric outputs; they should also present explanations that resonate with how people usually ask and answer "why" questions in daily life. This broader "human-centered" focus, Miller notes, calls for insights from the social sciences, where the processes of causal explanation, attribution, and sense-making in ordinary human interactions have been studied for decades.

Miller highlights that the resurgence of research in explainable AI (XAI) is response to ethical concerns and an accompanying lack of user trust. If an AI system denies a loan application or diagnoses a patient with a certain condition, relevant stakeholders (the applicant, the patient, regulatory bodies) often

[1] Or as the author puts it: " [..]*the solution to explainable AI is not just 'more AI.'*"

demand to know how those conclusions were reached. In outlining these constraints and requirements, Miller emphasizes that transparency is not a binary choice (completely open versus completely hidden). Rather, there are multiple facets to what makes an explanation "good," and these facets often involve psychological, social, or conversational dimensions. For example, people typically want concise reasons that address their specific query ("Why was my application declined rather than accepted?"), rather than the entire, exhaustive set of technical causes.

Building on this, studies from the social sciences reveal that people rarely want every possible contributing cause, but instead ask for an explanation in terms that are salient and intuitive within a particular context.[2] The significance for XAI is that AI designers can benefit from modeling explanations around these well-tested psychological patterns, thus producing interactions that feel more natural to end users.

2 Philosophical Foundations

The philosophical literature treats causation, not correlation, as the heart of any real explanation. Drawing on thinkers from Hume to Lewis, this section reframes explanations as answers to "why-questions" whose point is to reveal which factors actually made a difference and how replacing one factor with an alternative would have changed the outcome. Basing AI explanations on these estabished philosophical insights, we can move beyond mere pattern description toward explanations that genuinely satisfy human curiosity about causes and possibilities.

2.1 Philosophical Foundations and the Contrastive Nature of Explanation

Miller begins by positioning explanation within a long-standing philosophical and scientific tradition. He observes that although modern AI methodologies frequently emphasize statistical correlations and likelihood estimates, the philosophical literature typically emphasizes causal relationships as central to genuine explanations. A key insight from these philosophical perspectives is that merely highlighting correlations, for example, stating that "Feature A is associated with Outcome X", often does not address the deeper question of why something occurred. People usually seek to understand which specific factors genuinely caused an event, particularly how changing these factors might have altered the outcome.

In exploring this philosophical foundation, Miller highlights that explanations are not simply raw cause-and-effect sequences; rather, they represent a form of knowledge transfer designed specifically to answer a "why"-question. Even if a person or system could identify every factor that contributed to an

[2]See Major finding 2 in Section 1.2 and the discussion in Section 4 in [1]

outcome, it is generally impractical to convey the entire causal chain in everyday explanations, as doing so is usually too detailed or technically overwhelming. Philosophers such as David Lewis, and earlier David Hume, demonstrated that humans often reason about causation in a counterfactual manner: "If this particular factor had been absent, would the outcome still have occurred?" In essence, a factor is judged as causal if altering it, while holding all other conditions constant, would lead to a different outcome.

Two Major Philosophical Insights

Understanding what makes an account genuinely explanatory requires looking beyond the mere identification of causes. First, philosophers distinguish *causality* from explanation. While tracing the chain of events that produced an outcome is necessary, it is not in itself sufficient. A useful explanation also filters, organizes, and presents that causal information in a way people can readily grasp and apply. Even an exhaustive catalogue of physical or statistical processes must be carefully selected and framed before it becomes meaningful to a human audience.

Second, philosophical traditions emphasize that explanations are *inherently selective*. Most real-world events emerge from a tangle of contributing factors, yet listing every one of them is seldom helpful. In everyday contexts we typically want the most salient or illuminating causes, the handful of factors that convey why this event matters or how it differs from what we expected. Consequently, good explanations focus on the most relevant or noteworthy elements and leave the rest in the background, striking a balance between accuracy and clarity.

Together, these two insights show that explaining is an interpretive act: we must both uncover causes and decide which of them will best clarify the phenomenon for a particular audience and purpose.

Explanation as an Answer to "Why P instead of Q?"

One prominent theme Miller highlights from social sciences and philosophy is the inherently contrastive nature of explanations. When individuals ask, "Why did X happen?", they are typically comparing the event (implicitly or explicitly) to an alternative scenario. These alternative scenarios could include:

- Another real situation: "Why did the system give a low score here but a high score to another item?"
- A hypothetical or *counterfactual* scenario: "Why did the model reject my loan application (P) instead of approving it (Q)?"

Thus, when someone asks, "Why did it classify my image as a spider?", they usually have a particular alternative in mind—for example, "rather than a beetle" or "rather than no classification at all." Critically, addressing the question "Why?" properly requires identifying this implicit or explicit alternative scenario (the *foil*, Q, in "P rather than Q").

2.1.1 Why Contrast Matters

Events (such as classifications or recommendations) often involve multiple causal factors, so focusing on a specific contrast highlights the most salient ones.[3] For example, if a user asks, "Why is this image classified as a spider rather than a beetle?", the best explanation emphasizes the distinguishing features—like the number of legs, instead of listing all characteristics that define spiders. Psychologically, people seldom consider "Why P happened?" in isolation; instead, they evaluate competing hypotheses or alternatives (Q). Thus, an effective explanation addresses the specific contrast of interest, such as, "The image was classified as a spider because it has eight legs, whereas a beetle typically has six". Moreover, when someone simply asks, "Why P?" without explicitly mentioning a second scenario, Miller and others argue that there is always an implicit foil—"Why P rather than not-P?" or "Why P instead of some expected alternative?" This implicit foil often represents a default or "normal" situation. If the relevant foil is not made clear, the explanation may fail to address the real concern behind the question.

Recognizing the contrastive nature of explanations allows AI systems to deliver clearer, sharper responses by emphasizing the differences between a selected outcome and its relevant alternatives. By adopting a user-centric approach, AI systems can explicitly clarify the user's intended contrast, thereby resolving ambiguous queries and preventing information overload. Integrating such contrastive reasoning into AI explanations ensures that responses naturally align with human cognitive patterns, enhancing their clarity and effectiveness.

2.2 A Kantian Lens on Explanation and Autonomy

Immanuel Kant famously insisted that human knowledge is not a passive reflection of the world, but an active, structured process in which the mind imposes universal categories, most fundamentally space, time, and causality, onto experience.[4] Kant further argues that genuine explanation is inseparable from moral autonomy: humans explain actions not merely to describe what happened, but to justify choices against a backdrop of self-imposed, universal moral law [2]. Miller inherits this conviction, mainly that explanation must go beyond raw data, echoing the claim that people demand explanations that contrast a fact with a meaningful foil rather than reciting every causal detail. Like Kant, Miller treats explanation as inherently teleological, recognizing that humans ask "Why P rather than Q?" and expect answers framed in terms of purpose or intention. However, Miller departs sharply from Kant's normative, metaphysical project. Whereas Kant grounds explanation in the subject's transcendental categories and moral agency, Miller frames explanation as a descriptive, social-psychological process driven by cognitive biases (e.g. abnormality,

[3] Also referred to as Selective Focus

[4] "Space and time are the conditions under which alone we can have experience," he wrote, emphasizing that true understanding involves revealing the transcendental conditions that make phenomena intelligible. [2]

simplicity) and conversational norms. Miller explicitly rejects the notion that explanations must reveal ultimate, universal truths; instead, they must satisfy the explainee's epistemic needs, fostering trust and facilitating decision-making in context. In Kantian terms, AI has no noumenal access to the world or moral agency, it cannot legislate universal principles. Miller's XAI therefore seeks not to replicate Kantian autonomy, but to simulate its outward form by structuring machine-generated explanations in ways humans naturally understand.

Finally, Kant's account ties autonomy to the capacity for moral self-legislation, whereas Miller's explanatory autonomy is fundamentally instrumental: explanations empower users to interrogate, challenge, and refine an AI's output, but they do not imbue AI with genuine moral agency. In short, Miller inherits Kant's insight that explanation is neither passive nor purely mechanistic; but he replaces Kant's transcendental, normative grounding with a pragmatically oriented, empirically validated model of how people actually seek and evaluate explanations in everyday AI interactions.

3 Social Attribution and Intentionality

In this section of the paper, Miller turns from the broad philosophical grounding of explanation to a more specific strand of research—how people explain the behavior of other agents (especially other people, but also, by extension, anthropomorphized robots or AI systems). The focal question is: What do humans perceive as the "reasons" behind an action when they see an entity acting deliberately or, conversely, when they see behavior that appears accidental or unintentional? To answer that, Miller draws extensively on social psychology, particularly the foundational work of Fritz Heider and more recent contributions by Bertram Malle.

3.1 Heider's Legacy and the Birth of Social Attribution

3.1.1 Moving Shapes, Human Minds

A key stepping stone is Heider and Simmel's classic 1944 experiment, in which participants watched simple geometric shapes (a circle and two triangles) move around in a short video. Despite the shapes being just 2D animations, people spontaneously and inevitably described them with human-like intentions, motives, and even personality traits (e.g., "The big triangle is bullying the little triangle," "The circle wants to run away," etc.). From this, Heider famously concluded that humans are "wired" to interpret observable movements in intentional or goal-driven terms whenever there is even a hint that an entity is capable of acting on its own.

Heider took the idea further by distinguishing explanations of objects from explanations of agentic behavior. He insisted that people see intentional states (goals, beliefs, desires) behind human actions, and that these mental states often figure more prominently in explanations than purely mechanical or physical details. Thus, a physical event might be explained by forces and collisions,

but an intentional action is often explained in terms of a person's motives or knowledge states. Building on this insight, Miller underscores that people expect different kinds of explanations from AI systems depending on whether the system's behaviour looks intentional (e.g., a planning algorithm deliberately selecting among mission options) or incidental (e.g., a random glitch or sensor error).

3.1.2 Malle's Framework for Explaining Behavior

Building on Heider's seminal ideas, Bertram Malle developed a detailed conceptual framework known as the "folk-conceptual theory of explanation," which highlights how people parse human behavior into distinct explanatory types. First are *reason explanations*, which explicitly reference the mental states, beliefs, desires, intentions, that make sense of an action (e.g., "He grabbed an umbrella because he believed it might rain and he didn't want to get wet"). These are the everyday "He did it because he thought..." or "She did it because she wanted..." statements. Next come *causal histories of reasons*, which zoom out to consider the deeper background that shaped a person's beliefs or desires, such as upbringing, personality, or cultural norms ("He hates rain because his mother always fretted over wet clothes, so now he's extra cautious"). Unlike reason explanations, causal histories do not assume a direct, rational chain from cause to action; instead, they point to influences that predispose an agent to hold certain motivations. Finally, we have *enabling factors* which clarify how an action succeeded without necessarily addressing why it was chosen. An example might be "He managed to get home so fast because he borrowed a friend's car." The borrowed car is not the reason he decided to hurry home in the first place; it simply enabled him to carry out his plan more efficiently.

According to Malle, these distinctions matter when people try to understand AI systems. If an AI is viewed as deliberative—formulating goals, weighing options, and making choices—observers look for the AI's reasons: "What did it believe about the environment?" "What did it hope to accomplish?" If the AI fails or behaves unexpectedly, people may instead look for enabling factors (e.g., insufficient data or missing resources) or for broader causal histories (the way it was trained or coded) to explain why the AI had particular assumptions or tendencies. As Miller notes, recognizing which type of explanation users seek—whether they want the AI's immediate reason for an action, or an insight into its "personality" or long-term training history—helps ensure that the explanations provided match human expectations.

3.1.3 Intentional Behavior "Gets" Priority

A robust finding in social psychology is that people's explanations spotlight intentions and motives far more than physical or background factors when the event is perceived as deliberate. For instance, if a robotic system apparently "chose" to allocate resources in a certain way, the question is, "What was its goal or rationale?" rather than, "What is the mechanical path of causation inside its

circuit?" People prioritize understanding the mental or goal-driven aspect. On the other side, when something is unintentional (an accident or random glitch) human interpreters often revert to more mechanical or situational accounts. For example, "The sensor reading spiked, causing an unexpected system reaction." In moral or legal contexts, blame and responsibility also tie heavily to perceived intentionality. Miller uses these social attribution principles to suggest that; whenever AI systems exhibit complex, goal-directed behaviors, users will tend to attribute intentions, beliefs, and desires to them whether or not the AI literally has such constructs. Designers should anticipate these attributions and produce explanations aligned with the folk-psychological lens: for example, "The planner chose this route to minimize travel time," rather than simply stating, "The polynomial cost function had a lower value along that route."

Moreover, if the system's action violates user expectations or norms, people will further scrutinize the AI's intentions: "Did it do that on purpose?" They may look for deep causes, such as "It was programmed or trained in a way that doesn't factor in social preferences." By providing an explanation that addresses the AI's "motives," one aligns with how users naturally interpret behavior and thus can sustain trust and understanding.

4 Cognitive Biases in Explanation

Central Themes and Evidence

In this portion of the paper, Miller compiles and interprets a wide range of findings from cognitive psychology and social psychology about how people arrive at and assess explanations. While earlier sections focus on the structural aspects of explanation, here the focus shifts to the biases and heuristics that systematically shape which explanations humans find compelling, memorable, or satisfactory. These biases do not necessarily align with "objectively" correct statistical or causal accounts, but they play a defining role in practical, everyday explanation processes.

4.1 Abnormality, or the "Abnormal Conditions Focus Model"

One recurring theme is that people tend to highlight abnormal or unexpected factors when explaining an event, rather than mentioning routine or "normal" elements. Formally, this aligns with what some researchers call the abnormal conditions focus model, first proposed by legal scholars Hart and Honoré and later tested in psychological experiments. The core idea is that when we ask, "Why did this happen rather than something else?" we look for a factor that deviated from the usual or expected script. If everything about a situation is perfectly ordinary (the usual traffic, the usual route home), we typically do not zero in on those factors as "the cause," even if they were, in a purely physical sense, necessary for the outcome. Instead, we single out the one unusual factor, like a sudden road detour, a missing tool, or an unexpected phone call.

A case example Miller uses is the story of a fatal accident: the driver took an unusual scenic route, and there happened to be a teenage driver under the influence. Although many causes converged on this accident (the engine ran, the roads existed, the presence of oxygen, etc.), experiments show people fixate on the "unusual scenic route" or the "teen's drug use" as the cause. By emphasizing abnormal events, people effectively sift out any condition that appears routine or backgrounded. This aligns with a fundamental bias: we see ordinary factors as "given," and focus on the handful of elements that stand out from the baseline.

Closely tied to the idea of abnormality, Miller highlights two complementary processes, *discounting* and *backgrounding*, that shape how people either dismiss or push certain factors into the explanatory background. *Discounting* occurs when multiple plausible causes arise, but one is deemed more salient or significant than the others; for example, if someone is late to a meeting due to both oversleeping and a train delay, observers often focus on oversleeping (a factor within the person's control) and downplay the train delay. *Backgrounding*, on the other hand, happens when a condition is so common or widely shared across possible scenarios that it no longer stands out; if the train is always delayed, that no longer appears as an unusual factor worth highlighting. In practical terms for AI, if a system tries to account for every routine micro-cause, such as a charged laptop battery or stable Wi-Fi, alongside genuinely significant or surprising factors, it risks overwhelming users. Because people naturally background those ordinary conditions, an AI that fails to foreground what is truly out of the ordinary disrupts the user´s expectations for a clear and useful explanation.

4.2 Mutability and Counterfactual Thinking

A further cognitive tendency is mutability, which describes how readily people can imagine altering ("undoing") a particular event in the causal chain:

- **Temporality**: Studies show that people more often "undo" a recent event than a distal one. If a chain has multiple steps, they prefer blaming or changing the last step before the outcome.
- **Controllability and Intentionality**: Actions that someone chose (with freedom or intent) are deemed more "mutable" than purely accidental or environmental factors. As a result, an intentional action is singled out for blame or explanation more than an accidental factor, even if both contributed equally.
- **Social Norms**: If an event violates a norm (social, moral, or situational) people readily pick that violation as the cause. A classic example is the "Knobe effect".[5]

[5] In Knobe's original vignette, a corporate executive approves a new program knowing it will harm the environment as a side effect; most respondents judge the harm intentional. When the identical program is described as helping the environment, respondents say the benefit was unintentional.[3] Also see Section 3.5 in [1]

In all these cases, mutability is a form of mental simulation: "If we took away that factor, would the outcome have changed?" The psychologically "easiest" or most "salient" factor to mentally remove tends to end up as the prime cause in an explanation. One might assume that humans select explanations based on the most probable cause, but Miller presents evidence that people often favor simpler or more general accounts over strictly likeliest ones. This tendency resonates with a principle in cognitive psychology: that an explanation's *simplicity* and *generality* can outweigh raw probability. Explanations focusing on fewer key causes are seen as more coherent and plausible, while those unifying or clarifying multiple observations tend to be more satisfying, even if they are not the most statistically likely[6]. In addition, a purely statistical statement (e.g. "Feature A has a 0.87 correlation with outcome B") often feels less compelling than pointing to some mechanism or "why" behind the event. Consequently, a data-driven AI that simply presents the "most probable cause" could miss the mark. From a user's perspective, it is sometimes more persuasive to receive a simpler, more general explanation, such as "The system prioritizes cost-saving, so it picked the cheaper route," than a technically exhaustive breakdown of probabilities and correlations.

Another notable phenomenon is that in certain contexts, people commit the *conjunction fallacy*, judging a more specific explanation as more likely if it sounds more "representative" of known stereotypes. In other words, if cause A and cause B together form a coherent story, humans may observe them as more probable than a single cause A by itself, even though logically, $P(A) \geq P(A \wedge B)$. Miller points this out to illustrate that *coherence* or *narrative fit* can trump purely formal logic in the mind's search for best explanations.

4.3 Inherent vs. Extrinsic Features (Inherence Bias)

Lastly, Miller highlights how people tend to explain phenomena by pointing to *inherent* or essential traits rather than external accidents of circumstance. Called the *inherence bias*, this drives us to say, for example, "Spiders are scary because they have many legs," downplaying the fact that fear of spiders might be learned from a parent. This inclination to root explanations in "what the entity fundamentally is or does" can overshadow historical, cultural, or environmental causes. This matters for AI because when an intelligent system classifies images or diagnoses diseases, end users may search for *inherent* or essential features ("It labeled the image a spider because it's eight-legged") rather than the "extrinsic" factors like the distribution of the training data. Designers should consider that if the system's real reason depends on extrinsic, context-specific factors (such as subtle training anomalies), it may not align with how the user's mind "wants" to hear an explanation.

Ultimately, Miller's broader message is that these *cognitive biases* are not quirks

[6]Though truth or likelihood is still important, various experiments demonstrate that if a slightly less likely cause offers a more elegant account, people frequently gravitate toward it.

to be "corrected," but well-documented aspects of human reasoning that must be *accommodated* for explanations to be accepted and understood. Acknowledging cognitive biases thus helps bridge the gap between how an AI conceptualizes cause-and-effect[7] and how humans naturally reason about events. Recognizing these biases enables AI explanations to resonate with human expectations, emphasizing meaningful causal contrasts rather than exhaustive detail.

5 Explanation as Dialogue

In this section, Miller shifts from viewing explanations as one-shot statements of cause to understanding them as iterative, communicative activities, showcasing how explanation can also be dialogues, which are a interactive and social process. Drawing on Hilton, Antaki and Leudar, and Walton, he demonstrates how explanation unfolds through back-and-forth interaction between an explainer and an explainee, each guided by their own knowledge, expectations, and objectives.

5.1 Explanation Is Conversation, Not Just Attribution

Early AI research often reduced "explanation" to identifying and displaying the causes behind a system's outputs. While causal attribution is undoubtedly part of any explanation, Miller moves beyond causal attribution and stresses that focusing solely on cause overlooks how explanations are actually conveyed and interpreted between people. He draws on the influential work of H. P. Grice, who proposed four maxims of conversation that help ensure effective communication:

- **Quality**: Be truthful and support statements with evidence.
- **Quantity**: Provide the amount of information the situation requires, neither too much nor too little.
- **Relation (Relevance)**: Stay on-topic; address the actual question being asked.
- **Manner**: Communicate clearly, avoiding ambiguity and unnecessary complexity.

Adhering to these principles leads to more natural, helpful explanations than a mere "info dump." It also ensures that the explanation precisely targets the explainee's question, rather than overwhelming them with irrelevant details. Furthermore, in a dialogue, each participant has a mental model of what the other person knows and believes. Because explanations are more efficient when they address genuine knowledge gaps, an explainer should avoid reiterating known facts while supplying any crucial information the explainee lacks. This requires the explainer to continually assess the other person's epistemic state—what they

[7] Often via formal models, statistics, and probabilities

think the other person knows, doubts, or misunderstands—so they can tailor the explanation to match the explainee's current level of understanding. The reason for this is that an explanation is also only as good as its relevance to why the explainee is puzzled[8]. If an AI's decision surprises a user, perhaps because it conflicts with their prior assumptions, the explanation must tackle that specific point of confusion. Failing to address the user's underlying concern leaves them unsatisfied or mistrustful. Finally, because many explanations are contrastive in practice, the explainer often needs to elicit precisely which alternative the user is concerned about (e.g., "Are you wondering why it classified the image as a spider rather than a beetle, or rather than no insect at all?"). By clarifying this "foil," the explainer can focus on the relevant differences rather than offering a broad, unfocused response.

Follow-Up Questions and Requests for Justification

One significant idea is that real-world explanation rarely stops at a single statement. Rather, the explainee can respond with clarifications, counterexamples, or additional queries:

> "But how did you detect it had eight legs?"
> "Could it have been another eight-legged creature?"

This dialogic aspect lets the explainee probe deeper if they remain unconvinced or if the explanation triggers new questions. The explainer, in turn, can refine or expand the explanation. This iterative process also underlies explanatory debugging in machine learning contexts, where the user queries one cause, tests it, and then asks more questions if they still find the system's reasoning opaque.

5.1.1 Walton's Dialogue Models

Building on the work of philosopher Douglas Walton; Miller references a structured perspective in which explanations can be framed as a specific type of dialogue, a "clarification" or "explanatory" dialogue, distinct from others like negotiations or debates. Walton proposes:

- Opening Stage: Participants establish the purpose (the "why-question") and relevant assumptions.
- Exploration Stage: The explainer offers causal statements or reasons; the explainee can ask for evidence or clarifications.
- Closing Stage: The explainee either accepts they now understand, or transitions to a different dialogue (e.g., argumentation if they want to dispute the facts).

[8] Also referred to as "Epistemic Relevance"

This model draws parallels with everyday human conversation: we do not just provide a raw cause and walk away; we engage, check for understanding, and adapt. If a user remains confused or skeptical, the conversation may shift into a more argument-like exchange, where the explainer must defend or elaborate on points.

Furthermore, explanations convey far more than just factual information; they often hint at deeper motives and responsibilities. Research in discourse and social cognition shows that people routinely infer intent or bias.[9] For instance, if an AI's explanation highlights some causes while downplaying or omitting others, users may suspect the system (or its designers) of obscuring information or "spinning" the story. Moreover, if the explainee senses that the explainer has a personal or institutional stake in how the explanation is received, such as persuading them to trust the AI, they may dismiss the explanation or seek corroboration from independent sources. Recognizing these subtleties is crucial: even when the facts are technically correct, an explanation can fail if it overlooks the social dimension of trust and acceptance.

5.1.2 Designing AI for Dialogic Explanation

By treating explanation as a social conversation, Miller highlights several central insights[10]. First, explanations should not be treated as mere fact dumps; rather, they must be structured, relevant, and adaptable to the audience's ongoing questions. Second, Grice's cooperative principle[11] helps determine how much and what kind of information is appropriate at each turn. Furthermore, iterative refinement is the norm: the user asks an initial question, receives an explanation, and then follows up with new or more precise questions. Finally, understanding is co-created, depending on both the explainer's clarity and the user's willingness to engage. In this way, the user is not a passive observer but an active participant in shaping which causes are revealed, how they are framed, and when the conversation concludes. This shift, from viewing explanation as a simple "add-on" to recognizing it as an interactive, dialogic exchange, forms the foundation of Miller's human-centered approach to XAI, emphasizing collaboration, adaptability, and user involvement in the explanatory process.

6 Primary Contribution and Analysis of Related Approaches

The central contribution of this paper is to unite decades of philosophical, psychological, and linguistic research into a cohesive framework for explainable AI. Rather than viewing "explanation" as a mere technical exercise in exposing a model's internal workings, Miller contends that explanations must align with how humans naturally seek and evaluate causes. This means explanations

[9] An effect known as "implicature"

[10] See Appendix A for Miller's Dialogic-Explanation Design.

[11] See H.P. Grice's Cooperative Principle in the philosophy of language, section 5.1 or [4]

should be selective, contrastive, and contextual in ways that mirror everyday social interactions. In doing so, Miller offers a roadmap for designing systems that provide more than raw probabilities or exhaustive model traces; they deliver explanations that users naturally find trustworthy and insightful. Crucially, this synthesis of social science research offers AI practitioners concrete guidance for building user-centered explanatory tools that can facilitate learning, foster trust, and ultimately support more meaningful human-AI collaboration.

Related Approaches

The paper successfully demonstrates how effective explanations must align with the familiar human processes of causal attribution and conversational exchange: people do not simply want more raw data about model internals; they want explanations framed as answers to specific alternatives ("Why P rather than Q?") and tailored to their current concerns. On the other hand, Zachary C. Lipton's "The Mythos of Model Interpretability" critiques the very notion of interpretability as it is typically invoked in machine learning. Lipton argues that the community often uses "interpretability" as a catch-all term, conflating technical transparency, trust, fairness, and numerous other goals [5]. In doing so, researchers risk glossing over crucial distinctions between, for example, whether a model's parameters are easily inspected versus whether a user can truly grasp why the model behaves the way it does. Lipton shares Miller's skepticism about simplistic efforts (such as the mere display of model weights), but approaches the subject from a conceptual standpoint rather than a social-scientific one. Lipton calls for rigorous definitions of interpretability; Miller, on the other hand, prescribes an interdisciplinary approach grounded in how humans naturally construct and evaluate explanations. While both authors caution that "technical detail alone does not ensure real understanding", Lipton mostly interrogates the field's conceptual ambiguities, whereas Miller describes how to build explanations that speak to human social and cognitive biases.

Ribeiro, Singh, and Guestrin's work, "Why Should I Trust You? Explaining the Predictions of Any Classifier," offers a direct, algorithmic method called LIME (Local Interpretable Model-agnostic Explanations). Its goal is to generate local surrogate models for arbitrary black-box classifiers[12], spotlighting which features most strongly contribute to a specific output. Although Ribeiro et al. briefly discuss the importance of user trust, their solution is anchored in a more technical, retrospective approach. By perturbing the input around a single data instance and fitting a simple, linear model to those perturbations, LIME furnishes a concise picture of how the original classifier behaves in that local neighborhood [6]. This stands in contrast to Miller's broad, social-scientific lens: Miller advocates that explanations for AI must incorporate principles such as highlighting abnormal conditions, clarifying intentionality, and engaging in back-and-forth dialogue. LIME supplies with practical insight, but

[12] See [6] for the full algorithmic approach.

it does not delve into the richer psychological context of "why P rather than Q?" that Miller emphasizes. Indeed, Ribeiro et al. focus on providing a simplified feature-importance interface; Miller insists that any truly "human-friendly" explanation must account for the specific contrast a user cares about, the user's mental model, and how iterative questioning refines explanation.

While LIME illustrates the power of local surrogates, a complementary strand of XAI pursues global interpretability: decision-tree surrogates distilled from neural nets [7], symbolic distillation and model-based trees that approximate the full decision surface [8]. Global views[13] promise one stable story for all inputs, yet they abstract away those "Why P rather than Q?" contrasts that Miller argues users actually ask about. Miller's dialogic lens therefore problematizes a purely global stance: a one-shot tree may satisfy audit requirements, but it cannot negotiate follow-up "what-if" questions or reveal instance-level anomalies. Conversely, local surrogates such as LIME raise their own concern, stability. Small perturbations or different random seeds can yield diverging feature weights, imparing trust [9]. Choosing between global coherence and local faithfulness is thus not merely a technical trade-off; it also surfaces Miller's claim that explanations must remain interactive and contrast-sensitive to stay meaningful for humans.

Taken together, Lipton, Miller, and Ribeiro et al. occupy three different but mutually illuminating positions within the debate on explainable AI. Lipton's conceptual analysis exposes the definitional tangle around "interpretability." Ribeiro et al. propose an operational, algorithmic tool that works for any classifier. Miller, building on established theories in the social sciences, provides in-depth guidance on how to craft AI explanations that resonate with everyday human reasoning. By weaving psychological and philosophical research into AI, Miller's main contribution is to illustrate that a purely technical, data-centric view of "explanation" can fall short of what people want; explanations that reflect how humans generally pose and resolve "why" questions in the real world.

7 Discussion, Conclusion and Future Research Directions

One of Miller's key strengths lies in his interdisciplinary grounding, which provides a blueprint for AI explanations that closely mirror real-world human reasoning patterns. This alignment with human cognition stands in contrast to purely algorithmic approaches and helps practitioners see that simply providing more data or exposing neural weights rarely makes a system more interpretable

[13] "Global views" in the XAI literature are explanation methods that aim to capture how the entire model behaves across all possible inputs rather than explaining single predictions case-by-case.

to non-experts. Furthermore, Miller's strong emphasis on dialogue pushes the conversation beyond one-shot explanations. In doing so, he underscores how iterative clarification is often crucial for real understanding, an insight deeply relevant to user experience and trust-building in AI. These arguments are also highly adaptable across a broad range of applications. Whether one is developing an AI system for healthcare diagnostics or financial decision-making, the principles of selective, contrastive, and socially aware explanations hold true. This versatility, coupled with the synthesis of multiple strands of social-science research, makes his contribution highly beneficial to those seeking a more user-centered form of AI explainability.

Despite these strengths, Miller's vision poses certain practical challenges. For instance, moving from theoretical insights to scalable implementations can be daunting. While iterative, dialogue-based explanations align well with how humans process "why" questions, many real-world AI systems operate at massive scale, processing millions of transactions or decisions per day. Providing real-time, user-specific interactions that adapt to each individual's knowledge state or preferred contrast might be computationally demanding and organizationally complex.

Another gap relates to feasibility in high-stakes or legally regulated environments. In scenarios like healthcare or finance, explanations must often follow strict guidelines. Aligning these regulations with Miller's flexible, user-centric model could require extensive domain customization. Moreover, Miller's grounding in human social cognition may not fully account for cases where the end-users are themselves AI or software systems, situations in which formal proofs or auditable logs might be more critical than "human-like" conversation. Expanding on this, several scholars have extended Miller's arguments by examining how conversational explanations might integrate with techniques like counterfactual analysis and local surrogate models[6], [8]. Meanwhile, others question whether a purely social-scientific framework neglects deeper transparency aspects, such as direct interpretability of model components or explicit fairness metrics. For example, Lipton calls attention to how different stakeholders each need a specific type of interpretability [5], some want simpler models, others want thorough audits, and others just need proof that decisions are correct. Placing Miller's approach within this broader "interpretability ecosystem" highlights both its potential synergy with other methods and its relative weaknesses in areas like direct model introspection or large-scale integration into highly complex systems.

Future Research Directions: Frameworks for Scalable Explanations

A pressing research goal is to devise explanation frameworks that balance interactive, user-specific dialogue with computational efficiency. One promising avenue is the development of tiered explanation protocols: systems that open with a concise, high-level overview of a model's reasoning and let users drill

down into increasingly detailed, contrastive explanations only when needed. Exploring how to automate these layered explanations, perhaps through large language models that generate real-time responses while honoring domain constraints, offers a fertile area for experimentation. [10]

Miller provides a solid theoretical foundation, but empirical work is still needed to gauge how different explanation styles influence user satisfaction and comprehension. Important questions include whether spotlighting "abnormal conditions" in model outputs actually strengthens trust [11]; how domain specialists such as radiologists or financial auditors respond to contrastive explanations [12] compared with lay users; and which metrics best capture "understanding" and "trust" in real-time interactive systems. Carefully designed user studies, potentially combining eye-tracking, think-aloud protocols[14], and analysis of system logs [13], [14], will be crucial for testing whether Miller's insights consistently yield measurable gains in interpretability.

Interdisciplinary Collaborations

Finally, Miller's approach stands to benefit from integrating more deeply with cognitive psychology, neuroscience, and HCI. For example, neuroscientific research into how people process counterfactuals [15] or focus attention on "abnormal" events could guide the design of AI interfaces that highlight the most psychologically salient factors in a given explanation. Meanwhile, HCI experts could refine interactive explanation modules to ensure usability across diverse platforms (desktop, mobile, AR/VR).

- AI + Cognitive Psychology: Could real-time measurement of a user's cognitive load help an AI system modulate explanation detail?[14]
- AI + Neuroscience: Might insights into human causal reasoning pathways inform how we structure machine-generated explanations so they better match the brain's own inference processes? [15]
- AI + HCI: Studying how everyday users click through interactive explanation dashboards could reveal best practices for layout, pacing, and terminology.[16]

These interdisciplinary collaborations would enrich the practical effectiveness of Miller's ideas while broadening their theoretical underpinnings. Overall, while Miller's framework for user-centric, socially grounded AI explanations is compelling, it raises critical questions about scalability, regulation, and feasibility in real-world deployments. Addressing these questions calls for both innovative technical frameworks (to automate personalized, dialogue-based explanations) and rigorous empirical studies (to measure user comprehension and trust in varied contexts). By uniting expertise from social sciences, HCI, and

[14] **Think-aloud protocol**: a qualitative research method in which participants verbalize their thoughts in real time while performing a task, allowing investigators to capture the user's moment-to-moment reasoning and decision processes.

leading-edge ML methods; future efforts can clarify when, how, and why human-centered explanations best serve different audiences, and and how to harmonize these explanations with large-scale, highly complex AI systems.

Conclusion

In this thesis, the key research question asked how insights from the social sciences, particularly philosophy, psychology, and cognitive studies, can inform and improve explainable artificial intelligence (XAI). Specifically, we identified three core elements of human-centric explanation: contrastive approaches that answer "Why P rather than Q?", an emphasis on abnormal conditions that highlights unusual causal factors, and dialogue-based methods supporting iterative user clarification. Drawing on foundational work in attribution theory, cognitive biases, and conversational pragmatics, we showed that people generally prefer explanations rooted in concrete human motivations and uncommon events over purely statistical or data-driven accounts. This socially informed perspective offers guidance for building AI systems capable of giving more intuitive, trust-enhancing justifications.

Looking ahead, socially grounded XAI approaches stand to reshape ethical, regulatory, and practical considerations in domains such as healthcare, finance, and autonomous systems. Although the discussion underscored challenges like scalability, privacy requirements, and domain-specific regulations, these same hurdles present opportunities for empirical testing, user-centered prototyping, and multidisciplinary collaborations. By involving stakeholders in iterative design studies, where users can interact with prototype explanation tools and provide feedback, AI practitioners can refine explanation styles without sacrificing computational efficiency or legal compliance.

In closing, adopting a human-centered, dialogue-driven strategy in AI explanation can help close the divide between sophisticated, opaque models and the communities that rely on them for critical decisions. Recognizing how people naturally seek explanations prompts the development of systems that are not just more transparent, but also more trusted and ethically defensible. As XAI research continues to evolve, embracing these psychological and social realities will be vital for innovations that align intelligent technologies more closely with human values.

References

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370218305988

[2] I. Kant, *Critique of Pure Reason*. Cambridge: Cambridge University Press, 1998.

[3] J. Knobe, "Intentional action and side effects in ordinary language," *Analysis*, vol. 63, no. 3, pp. 190–194, 2003.

[4] H. P. Grice, "Logic and conversation," in *Syntax and Semantics, Vol. 3: Speech Acts*, P. Cole and J. L. Morgan, Eds. New York: Academic Press, 1975, pp. 41–58.

[5] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018. [Online]. Available: https://arxiv.org/abs/1606.03490

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, 2016, pp. 1135–1144. [Online]. Available: https://arxiv.org/abs/1602.04938

[7] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," in *arXiv:1711.09784*, 2017. [Online]. Available: https://arxiv.org/abs/1711.09784

[8] J. Herbinger, S. Dandl, F. K. Ewald, S. Loibl, and G. Casalicchio, "Leveraging model-based trees as interpretable surrogate models for model distillation," 2023. [Online]. Available: https://arxiv.org/abs/2310.03112

[9] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021. [Online]. Available: https://www.mdpi.com/2504-4990/3/3/27

[10] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *arXiv preprint arXiv:2309.01029*, 2023, v3, 28 Nov 2023. [Online]. Available: https://arxiv.org/abs/2309.01029

[11] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: A user study," in *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing

Machinery, 2020, pp. 1–11. [Online]. Available: https://doi.org/10.1145/3377325.3377519

[12] P. Tschandl, C. Rosendahl, H. Kittler *et al.*, "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234, 2020. [Online]. Available: https://doi.org/10.1038/s41591-020-0942-0

[13] T. Kulesza, S. Amershi, M. C. Fifield, I. Popescu, and M. Burnett, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. Association for Computing Machinery, 2015, pp. 126–137. [Online]. Available: https://doi.org/10.1145/2678025.2701399

[14] L. Herm, "Impact of explainable ai on cognitive load: Insights from an empirical study," in *Proceedings of the 31st European Conference on Information Systems (ECIS 2023)*, Kristiansand, Norway, 2023, research Paper. [Online]. Available: https://aisel.aisnet.org/ecis2023_rp/269

[15] S. Huang, L. Faul, N. Parikh, K. S. LaBar, and F. De Brigard, "Counterfactual thinking induces different neural patterns of memory modification in anxious individuals," *Scientific Reports*, vol. 14, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-61545-x

[16] G. He, N. Aishwarya, and U. Gadiraju, "Is conversational xai all you need? human–ai decision making with a conversational xai assistant," in *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, 2025, pp. 907–924. [Online]. Available: https://dl.acm.org/doi/10.1145/3708359.3712133

A Appendix: Dialogic-Explanation Design

According to Miller, truly "explainable" AI systems should incorporate the idea of *interactive, iterative explanation*. Concretely:

1. **Context Awareness**
   - Where the system might keep track of what the user has asked so far and the user's apparent knowledge level.
   - Subsequent explanations adapt to that context, avoiding repetition and focusing on the user's points of confusion.
2. **Mixed Modalities**
   - Explanations need not be purely textual; they might include highlights on an image, visual flows through a plan, or step-by-step timelines.
3. **User Control over Depth**
   - The system can present a brief explanation first, then let the user "drill down" by posing follow-up questions. This approach aligns with the Gricean maxims, preventing information overload at the outset.
4. **Argumentation Capability**
   - If the user challenges the explanation, the system can provide additional justification, citing data or causal rules. This allows the user to *test* or *probe* the explanation rather than only passively receiving it.