Journal of Machine Learning Research XX (2025) 1–XX Submitted 7/25; Published

Are All Explanations Explainable? A Gödel-Like Theorem for Machine Learning

Someindra Kumar Singh SOMEINDRAS@GMAIL.COM
India

Editor: To be assigned

Abstract

As machine learning systems become increasingly complex and widely deployed, the demand for interpretable and trustworthy explanations has intensified. While numerous explanation methods have been proposed, most rely on intuitive desiderata rather than formal reasoning. In this work, we introduce an axiomatic framework for explainability in machine learning by defining a formal system that captures explanation logic through a set of well-motivated axioms. Our approach allows for reasoning about which explanations are derivable within the system and, crucially, whether certain valid explanations may be fundamentally unprovable.

Drawing inspiration from Gödel's incompleteness theorems, we prove that, under plausible assumptions, there exist truths about model behavior that cannot be formally explained within the axiomatic system itself. We introduce Gödel numbering to encode explanation statements as numbers and construct self-referential explanation formulas. We refine epistemic axioms to reflect practical limits on user knowledge and condition inference rules on model properties such as additivity and feature independence. This establishes a foundational boundary on the scope of algorithmic interpretability and invites the development of meta-explanatory systems beyond first-order frameworks.

Keywords: axiomatic explanation, causality, epistemic logic, explainability, formal methods, Gödel incompleteness, interpretability, logic, machine learning, unprovability

1 Introduction

As machine learning (ML) models are increasingly deployed in high-stakes domains such as healthcare, law, and finance, the demand for explainable and interpretable decisions has become critical. Stakeholders ranging from policymakers to practitioners seek not only accurate predictions, but also meaningful justifications for model outputs. This has led to a proliferation of explanation methods—such as SHAP, LIME, saliency maps, and counterfactual reasoning—aimed at shedding light on complex black-box models.

Despite their popularity, most explanation techniques are grounded in informal principles and lack a unified theoretical foundation. Existing approaches often rely on post hoc heuristics or intuitive desiderata without formally characterizing the logical structure of explanations or their limits. As a result, we have little understanding of whether such systems can, in principle, produce all valid explanations—or whether there exist inherent boundaries to what can be explained algorithmically.

This work draws inspiration from Gödel's incompleteness theorems in formal logic, which revealed fundamental limitations in provability within any sufficiently expressive axiomatic

system. In a similar spirit, we ask: *Are there limitations to what can be explained about machine learning models within a given formal framework?* Put differently, can we identify the boundaries of explainability itself?

To address this, we introduce a formal logical system $\mathcal{E}$ designed to reason about explanations in machine learning. We define a formal language, a set of axioms grounded in principles from the explainable AI (XAI) literature, and a proof system for deriving explanations. We then explore the metatheoretical properties of $\mathcal{E}$, showing that—even under sound and intuitive axioms—there may exist valid explanations that cannot be derived within the system. This mirrors Gödel's insight, applied to the domain of ML interpretability.

Our contributions are as follows:

- We introduce a formal logical system $\mathcal{E}$ for modeling machine learning explanations.
- We propose a set of axioms that encode widely accepted principles in XAI, such as additivity, sensitivity, and minimality.
- We refine the epistemic reasoning component to reflect realistic limits on user knowledge using modal logic.
- We formalize Gödel numbering within the language $\mathcal{L}_E$, enabling the construction of self-referential explanation formulas.
- We condition decomposition and aggregation inference rules on structural assumptions such as additivity and interaction-freeness.
- We prove a Gödel-like incompleteness result, showing that there exist true explanations that cannot be derived within $\mathcal{E}$.

This work lays a mathematical foundation for future research into the provability, completeness,
and epistemic boundaries of interpretability in machine learning.

2 Background and Related Work

2.1 Axiomatic Approaches in Explainable AI

In recent years, explainable AI (XAI) methods have increasingly adopted axiomatic frameworks to justify their design. Notably, Sundararajan et al. Sundararajan et al. (2017) introduced *Integrated Gradients*, an attribution method based on two key axioms: *sensitivity* and *implementation invariance*. Similarly, SHAP (SHapley Additive exPlanations) Lundberg and Lee (2017) builds upon the principles of Shapley values from cooperative game theory, adhering to axioms like *additivity*, *symmetry*, and *dummy*. These axioms are intended to define what a "good" explanation should look like and help justify the methods' correctness.

While these approaches have established important standards for evaluating and designing explanation techniques, they typically focus on justifying a specific method rather than constructing a general, logic-based foundation for explanations themselves. Moreover, these axioms are often used as design heuristics or evaluation criteria rather than components of a formal deductive system in which one can prove or disprove the existence of explanations.

2.2 Logical Foundations and Gödel's Incompleteness Theorems

Our work draws conceptual inspiration from the field of mathematical logic, particularly Gödel's incompleteness theorems Gödel (1931). Gödel famously showed that in any sufficiently expressive, consistent, and recursively enumerable formal system, there exist true statements that cannot be derived within the system. This result exposed fundamental limitations of formal reasoning and had profound implications across mathematics, philosophy, and computer science.

We view explanation in machine learning as a form of inference—attempting to derive rationales from model behavior. This perspective opens the door to treating explainability itself as a subject of formal investigation. Just as Gödel asked what truths are provable in mathematics, we ask: *What explanations are derivable in a formal system of XAI? And do there exist explanations that are true but inherently unprovable within any such system?*

2.3 Gap in Existing Work

Although axioms have been used to motivate or justify explanation methods, there has been no comprehensive effort to construct a general-purpose formal system for reasoning about explainability. Current work stops at defining or verifying specific properties (e.g., faithfulness, completeness, or minimality) for individual explanation techniques, without addressing the meta-theoretical question of whether explanation systems themselves are complete.

Our work fills this gap by:

- Introducing a formal logical system $\mathcal{E}$ for modeling and reasoning about explanations;
- Proposing axioms derived from well-established XAI principles;
- Demonstrating inference procedures and formal derivations;
- Proving that some explanations are logically unprovable within $\mathcal{E}$, thereby establishing foundational limits of interpretability.

This shift—from evaluating explanations to understanding the limits of explanation systems themselves—marks a new direction in the theoretical foundations of explainable machine learning.

See also Halpern and Pearl (2005); Pearl and Mackenzie (2019); Sundararajan et al. (2017); Bringsjord et al. (2021); Samek et al. (2021); Doshi-Velez and Kim (2017).

3 Formal Language for Explanations

To reason about explainability in a formal setting, we introduce a logical language $\mathcal{L}_{\mathcal{E}}$ that allows us to represent explanations, features, models, and their relationships. This language serves as the foundation for the axiomatic system $\mathcal{E}$ developed in later sections.

3.1 Syntax

The syntax of $\mathcal{L}_{\mathcal{E}}$ includes the following components:

Constants.

- $f, f_1, f_2, \ldots \in \mathcal{F}$: symbols representing machine learning models.
- $x, x_1, x_2, \ldots \in \mathcal{X}$: input instances.
- $y, y_1, y_2, \ldots \in \mathcal{Y}$: model outputs.
- $\phi_1, \phi_2, \ldots \in \Phi$: individual features or feature indices.
- $a_\phi \in \mathbb{R}$: real-valued attributions assigned to features.

Logical Connectives and Quantifiers. Standard first-order logic connectives and quantifiers:

$$\neg, \wedge, \vee, \rightarrow, \leftrightarrow, \forall, \exists$$

Explanation Predicates. We define several predicates to capture different aspects of explainability:

- Explains(f, x, ϕ, a): Feature ϕ receives attribution a as an explanation for $f(x)$.
- Explains(f, x, S): Subset of features $S \subseteq \Phi$ jointly explain the model output for input x.
- Contributes(ϕ, f, x): Feature ϕ causally influences the output of f on input x.
- Minimal(S): Set S is a minimal explanation—no proper subset of S also explains the output.
- Faithful(f, x, ϕ, a): Attribution a accurately reflects the local behavior of f around x.

3.2 Semantics

We interpret formulas in $\mathcal{L}_\mathcal{E}$ over models that represent machine learning systems, datasets, and attribution mechanisms. Semantics can be given in two forms: model-theoretic and probabilistic.

Model-Theoretic Semantics. A model $\mathcal{M}$ assigns meanings to the symbols:

- $f(x)$ denotes the actual output of model f on input x.
- Explains(f, x, ϕ, a) is true in $\mathcal{M}$ if an explanation system assigns attribution a to feature ϕ for input x.
- Contributes(ϕ, f, x) is true if interventions on ϕ can causally change $f(x)$.

Probabilistic Semantics. Alternatively, if the model f is stochastic or the data is probabilistic:

- Attributions are interpreted as expected contributions: $a_\phi = \mathbb{E}_x[\Delta f(x) \mid \phi]$.
- Faithful(f, x, ϕ, a) holds if the attribution a approximates the expected local gradient or Shapley value.

Optional: Modal and Epistemic Extensions. We may also enrich the language with modal operators to represent epistemic properties:

- $\Box\,\text{Explains}(f, x, \phi, a)$: "It is necessarily the case that ϕ explains $f(x)$ with attribution a".
- $\mathcal{K}_{\text{user}}\,\text{Explains}(f, x, \phi, a)$: "The user knows ϕ explains $f(x)$ with attribution a".

These extensions allow the system to capture not only what is true, but what is knowable or justifiable—important distinctions in explainability contexts.

4 Axiomatic System $\mathcal{E}$

We now define an axiomatic system $\mathcal{E}$ for reasoning about explanations in machine learning. This system formalizes intuitive principles commonly found in explainability methods (such as SHAP or Integrated Gradients) and enables formal derivation of explanation statements. The system consists of a set of well-defined axioms, together with rules of inference, forming the basis for provability and formal reasoning about explanations.

4.1 Core Axioms

The following axioms define fundamental properties that valid explanations should satisfy. These are grounded in the literature but here formalized for use in a logical system.

Axiom A1 (Additivity / Completeness).

$$\sum_{\phi_i \in \Phi} a_{\phi_i}(x) = f(x) - f(x_0)$$

For a given model f, input x, and baseline input x_0, the sum of individual feature attributions equals the difference in output.

Axiom A2 (Sensitivity).

$$\left(\exists x', \phi_i \text{ s.t. } f(x) \neq f(x') \wedge x_j = x'_j \ \forall j \neq i\right) \Rightarrow a_{\phi_i}(x) \neq 0$$

If changing feature ϕ_i alone causes a change in the model output, then its attribution must be non-zero.

Axiom A3 (Symmetry / Anonymity).

$$(\forall x,\ f(x) = f(\pi(x))) \Rightarrow a_{\phi_i}(x) = a_{\phi_{\pi(i)}}(x)$$

If two features are functionally identical under some permutation π, then they should receive equal attribution.

Axiom A4 (Local Consistency).

$$\text{If Contributes}(\phi_i, f', x) > \text{Contributes}(\phi_i, f, x),$$
$$\text{then} \quad a^{f'}_{\phi_i}(x) \geq a^{f}_{\phi_i}(x)$$

If a feature contributes more in a new model f', then its attribution should not decrease.

Axiom A5 (Minimality).

$$\text{Explains}(f, x, S) \wedge \exists S' \subset S \text{ such that } \text{Explains}(f, x, S') \Rightarrow \neg\text{Minimal}(S)$$

A set S is not minimal if a proper subset also serves as a valid explanation.

4.2 Optional Axioms

Depending on the context, additional axioms may be introduced, such as:

Axiom A6 (Determinism).

$$f \text{ and } x \text{ are deterministic} \Rightarrow \text{Explains}(f, x, S) \text{ is uniquely defined}$$

Axiom A7 (Epistemic Possibility)

$$\text{Explains}(f, x, \varphi, a) \Rightarrow \Diamond K_{\text{user}}\text{Explains}(f, x, \varphi, a)$$

If an explanation is valid, then it is *possibly* knowable to the user. This weakens the original epistemic assumption by avoiding the claim that all valid explanations are necessarily accessible to users. It acknowledges real-world limitations such as model complexity, lack of transparency, or cognitive constraints that may render certain explanations effectively inaccessible despite being semantically valid.

5 Gödel Numbering and Meta-Encoding

To enable self-referential reasoning within the explanation language $\mathcal{L}_E$, we define a Gödel numbering scheme that encodes every syntactic element of the system as a unique natural number.

Let $\mathcal{A}$ denote the finite alphabet of symbols in $\mathcal{L}_E$, including:

- Constants: f, x, φ, a, S
- Logical connectives: $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$
- Quantifiers: $\forall, \exists$
- Predicates: $\text{Explains}, \text{Contributes}, \text{Minimal}, \text{Faithful}, \ldots$
- Parentheses and punctuation

We define a mapping $g : \mathcal{A}^* \rightarrow \mathbb{N}$ such that each symbol is assigned a unique prime-based code, and full formulas are encoded via prime exponentiation. For example:

$$g(\text{Explains}(f, x, \varphi, a)) = 2^{c_1} \cdot 3^{c_2} \cdot 5^{c_3} \cdots$$

where each c_i corresponds to the code of the i-th symbol in the formula string.

Similarly, derivations (proof sequences) can be encoded as finite sequences of Gödel numbers. This allows us to define predicates such as:

- $\texttt{Proof}(p, \varphi)$: "$p$ is a valid proof of formula φ"
- $\texttt{Provable}(\varphi) \equiv \exists p\, \texttt{Proof}(p, \varphi)$

Because these predicates can themselves be expressed in $\mathcal{L}_E$, the system becomes self-referentially expressive. This enables the construction of fixed-point formulas such as:

$$\varphi_G := \neg\texttt{Provable}(\ulcorner \varphi_G \urcorner)$$

where $\ulcorner \varphi_G \urcorner$ denotes the Gödel number of φ_G.

This machinery forms the basis for the incompleteness result presented in Section 6, demonstrating that some semantically valid explanation statements are unprovable within the system $\mathcal{E}$.

6 Inference Rules and Derivation System

To reason formally within the axiomatic system $\mathcal{E}$, we define a set of inference rules that govern how new explanation statements can be derived from existing ones. Together, the axioms of Section 4 and the inference rules in this section form a complete deductive system to prove the properties of explanations in machine learning.

6.1 Derivability and Proofs

A statement φ is said to be *derivable* in $\mathcal{E}$, written $\mathcal{E} \vdash \varphi$, if there exists a finite sequence of applications of axioms and inference rules that yields φ. A *proof* in $\mathcal{E}$ is such a sequence.

6.2 Core Inference Rules

(MP) Modus Ponens.

$$\frac{\varphi \to \psi \quad \varphi}{\psi}$$

If $\varphi \to \psi$ and φ are both derivable, then ψ is also derivable.

(GEN) Universal Generalization.

$$\frac{\varphi(x)}{\forall x\, \varphi(x)}$$

If $\varphi(x)$ holds for arbitrary x, then it holds for all x.

(SUB) Explanation Substitution:

$$\forall f\, \forall x\, \forall a \left[\text{Explains}(f, x, \varphi, a) \land \varphi = \varphi' \Rightarrow \text{Explains}(f, x, \varphi', a)\right]$$

If two features φ and φ' are equal, then their attributions are interchangeable in any explanation involving the same model f, input x, and value a.

(DECOMP) Explanation Decomposition (Additive Models Only):

$$\begin{aligned}&\text{Additive}(f, x) \wedge \ \text{Explains}(f, x, S)\\ &\Rightarrow \ \forall\, \varphi_i \in S, \ \text{Explains}(f, x, \varphi_i, a_{\varphi_i})\end{aligned}$$

(AGGR) Attribution Aggregation (No Interaction Effects):

$$\begin{aligned}&(\forall \varphi_i \in S, \text{Explains}(f, x, \varphi_i, a_{\varphi_i}))\\ &\wedge \text{InteractionFree}(S, f, x)\\ &\Rightarrow \text{Explains}(f, x, S)\end{aligned}$$

If all features in S individually explain the output, and there are no interactions among them at x, then they can be aggregated into a joint explanation.

6.3 Soundness and Completeness (Preliminaries)

We define the system $\mathcal{E}$ to be:

- **Sound** if all derivable statements are semantically valid.
- **Complete** if all semantically valid statements are derivable.

In Section 6, we will show that under mild assumptions, the system $\mathcal{E}$ is sound but not complete. That is, there exist valid explanation statements that are true in all models of the system but cannot be derived using any finite sequence of axioms and inference rules. This establishes a formal limit on explainability analogous to Gödel's incompleteness theorems.

7 Gödelian Limits of Explainability

The axiomatic system $\mathcal{E}$ allows formal reasoning about explanations in machine learning using principles grounded in widely accepted XAI methodologies. However, just as Gödel's incompleteness theorems showed that not all true statements are provable in sufficiently rich formal systems, we now show that not all valid explanations are provable within $\mathcal{E}$.

7.1 Motivation and Intuition

Gödel's first incompleteness theorem states that any consistent, recursively enumerable, and sufficiently expressive formal system cannot prove all truths expressible within its own language. This idea has analogs in computational learning theory (e.g., the no free lunch theorems) and algorithmic information theory (e.g., Kolmogorov complexity).

In our context, the key question is: *Are there valid model explanations that, while semantically correct, cannot be derived using any finite sequence of axioms and inference rules from $\mathcal{E}$?* We argue that under natural assumptions, the answer is yes.

7.2 Assumptions

Let us assume:

- $\mathcal{E}$ is consistent: it does not derive contradictions.

- $\mathcal{E}$ is recursively enumerable: all proofs can be enumerated by a Turing machine.
- $\mathcal{E}$ is expressive enough to represent statements about its own derivations (e.g., via Gödel numbering or meta-encoding).

These are standard in logic and are satisfied by most formal systems used in mathematics or logic-based AI.

7.3 Self-Referential Explanation Construction

We define the self-referential explanation formula E_G as follows:

$$E_G := \neg \texttt{Provable}(\ulcorner E_G \urcorner)$$

Here, $\ulcorner E_G \urcorner$ denotes the Gödel number of the formula E_G, and the predicate $\texttt{Provable}(n)$ asserts that there exists a valid derivation of the formula with code n in the system $\mathcal{E}$. This construction leverages the fixed-point theorem enabled by the Gödel encoding machinery introduced in Section 5.

Let E_G be the following statement in the language $\mathcal{L}_{\mathcal{E}}$:

$$E_G := \text{“There is no derivation of } E_G \text{ within } \mathcal{E}''.$$

That is, E_G encodes its own unexplainability. If $\mathcal{E}$ were able to derive E_G, it would be inconsistent (since it would derive that E_G is unprovable while deriving it). If it cannot derive E_G, then E_G is true (assuming consistency), but unprovable—i.e., a valid explanation that cannot be derived in $\mathcal{E}$.

This mirrors Gödel's original proof, but in the context of explainability.

7.4 Theorem (Incompleteness of Explanation Systems)

Theorem: If $\mathcal{E}$ is consistent, recursively enumerable, and expressive enough to encode its own derivability, then there exists a valid explanation statement $E_G \in \mathcal{L}_{\mathcal{E}}$ such that:

$$\text{Semantically valid: } \models E_G \quad \text{but} \quad \mathcal{E} \nvdash E_G.$$

Proof Sketch: Construct E_G as a self-referential statement asserting its own unprovability. The formal machinery is similar to Gödel numbering and fixed-point theorems in Peano Arithmetic. Assuming consistency, E_G must be true, yet not derivable within the system.

7.5 Implications

This result implies that even in rigorously defined explanation systems, there will always exist model behaviors or justifications that cannot be formally derived. These unprovable explanations may nonetheless be intuitively or empirically valid.

This has several important consequences:

- No explanation system can be both complete and consistent if it is expressive enough.
- Explanation systems may require meta-explanatory tools or external reasoning layers.

- Practical explainability algorithms must accept epistemic limits on what they can justify.

In the next section, we reflect on the significance of this incompleteness result and explore potential directions for developing meta-systems that reason about or extend $\mathcal{E}$.

8 Probabilistic and Causal Extensions

While the base system $\mathcal{E}$ models explanations in a logical and deterministic setting, many real-world models operate under uncertainty and are sensitive to interventions. To handle such cases, we introduce probabilistic and causal semantics into the framework, extending its expressiveness to handle uncertain and counterfactual explanations.

8.1 Probabilistic Semantics

Let $\mathbb{P}(x)$ denote a probability distribution over inputs, and let $f(x)$ be a random variable induced by stochastic components in the model or input.

Expected Attribution. We redefine the attribution of a feature ϕ_i at input x as the expected marginal contribution:

$$a_{\phi_i}(x) = \mathbb{E}_{S \subseteq \Phi \setminus \{\phi_i\}} \left[f(x_S \cup \phi_i) - f(x_S) \right],$$

where x_S denotes a sample with only features in S fixed and others marginalized.

This corresponds to the Shapley value from cooperative game theory and is compatible with the additive axiom under probabilistic settings.

Probabilistic Explanation Predicate. We define:

$$\text{Explains}_{\mathbb{P}}(f, x, \phi, a) := \mathbb{E}[a_\phi(x)] = a,$$

indicating that on average, ϕ contributes amount a to the model output.

8.2 Causal Semantics

We introduce a causal model $\mathcal{M} = (\mathcal{U}, \mathcal{V}, F)$, where:

- $\mathcal{U}$: exogenous variables,
- $\mathcal{V}$: observed variables including features and output,
- F: structural equations.

Let $\text{do}(\phi_i = v)$ represent an intervention setting feature ϕ_i to value v.

Causal Attribution. The causal effect of feature ϕ_i on output Y under input x is:

$$\text{CE}(\phi_i, x) = \mathbb{E}[Y \mid \text{do}(\phi_i = x_{\phi_i})] - \mathbb{E}[Y \mid \text{do}(\phi_i = x'_{\phi_i})]$$

for some reference value x'_{ϕ_i}.

We define a causal explanation predicate:

$$\text{Contributes}_{\text{causal}}(\phi_i, f, x) := \text{CE}(\phi_i, x) \neq 0$$

Causal Minimality Axiom.

$$\text{CausalMinimal}(S) := \neg\exists T \subset S \text{ such that } \text{Explains}(f, x, T)$$

$$\text{and} \quad \forall \varphi_j \in S, \text{CE}(\varphi_j, x) \neq 0$$

8.3 Probabilistic Soundness

We say $\mathcal{E}_{\mathbb{P}}$ is probabilistically sound if:

$$\mathcal{E}_{\mathbb{P}} \vdash \varphi \Rightarrow \mathbb{P} \models \varphi$$

i.e., all derivable explanation statements hold in expectation under the data distribution.

8.4 Counterfactual Reasoning

We can also reason about *what would have happened* had certain features been different. Define:

$$\text{CounterfactualExplains}(f, x, \phi_i, x') := f(x) \neq f(x_{\setminus \phi_i} \cup x'_{\phi_i})$$

meaning that changing ϕ_i to its value in x' changes the output — a direct counterfactual test for explanation.

This enables extending $\mathcal{E}$ with modal reasoning for explanations based on potential outcomes.

9 Proof Search and Automation

The axiomatic system $\mathcal{E}$ enables formal derivation of explanation statements. To apply this system in practice, we introduce an algorithmic framework for automated proof search. This allows us to verify (or refute) whether a given explanation follows logically from the axioms and inference rules, using a bounded reasoning procedure.

9.1 Problem Definition

Let $\mathcal{E} = (\mathcal{A}, \mathcal{R})$ be the explanation system, where:

- $\mathcal{A}$ is the set of axioms,
- $\mathcal{R}$ is the set of inference rules (e.g., Modus Ponens, Attribution Aggregation).

Given a target formula φ (e.g., $\text{Explains}(f, x, \phi, a)$), we define:
Goal: Determine whether $\mathcal{E} \vdash \varphi$, i.e., whether φ is provable within the system.

9.2 High-Level Strategy

We implement a bounded-depth backward proof search that attempts to construct a derivation tree for φ from the axioms and inference rules. If no derivation is found within a given depth or time bound, we return *undecided*.

9.3 Proof Search Algorithm

Algorithm 1 ExplanationProofSearch(φ, $\mathcal{E}$, depth)

Require: Explanation system $\mathcal{E} = (\mathcal{A}, \mathcal{R})$, target formula φ, depth limit d

Initialize agenda: $Q \leftarrow \{(\varphi, 0)\}$
Initialize proof trace: $T \leftarrow \emptyset$
while Q not empty **do**
  Pop (ψ, k) from Q
  if $\psi \in \mathcal{A}$ **then**
    Add ψ to trace T
    continue
  end if
  if $k \geq d$ **then**
    return "Undecided (depth limit reached)"
  end if
  for all rule $r \in \mathcal{R}$ **do**
    Use syntactic unification to find premises $\psi_1, \ldots, \psi_n$ such that applying r derives ψ
    for all premise ψ_i **do**
      Add $(\psi_i, k+1)$ to Q
    end for
    Record $\psi \leftarrow (\psi_1, \ldots, \psi_n)$ in trace T
  end for
end while
if every $\psi_i \in T$ is either an axiom or derived by a rule in $\mathcal{R}$ **then**
  return "Proof found", trace T
else
  return "Not provable within bounds"
end if

9.4 Interpretation

The algorithm performs backward chaining from the goal, attempting to apply inference rules in reverse until reaching axioms. A successful derivation yields a valid explanation proof trace. Failure to derive within bounds indicates either:

- φ is unprovable in $\mathcal{E}$, or
- More inference depth or rules are required.

9.5 Applications

- **Explanation Verification:** Automatically verify whether a proposed attribution or explanation satisfies the axioms and can be formally derived.
- **Refutation and Counterexamples:** If proof search fails, the system may generate counterexamples (e.g., specific model-input pairs) where axioms don't hold.

- **Debugging Explanation Methods:** Use proof failure traces to diagnose which axiom or inference failed, helping refine or repair explanation algorithms.

- **Automation in Theorem Provers:** The formal language $\mathcal{L}_{\mathcal{E}}$ can be encoded in SMT or HOL provers (e.g., Z3, Coq, Lean), allowing integration into formal verification pipelines.

10 Discussion

Our incompleteness result establishes a fundamental limitation on the power of any formal explanation system. Just as Gödel's theorem revealed that no axiomatic system for arithmetic can capture all truths about numbers, we show that no sufficiently expressive and consistent system for explainability can capture all valid explanations about machine learning models.

10.1 Interpretability Beyond Formal Derivation

In practice, explanation methods are often evaluated based on human interpretability, stability, or alignment with domain intuition. Our result suggests that some intuitively correct explanations may fall outside the scope of any fixed deductive system. This highlights the need for more flexible, layered approaches to interpretability that go beyond first-order logic.

10.2 Limits of Axiomatic Justification

Many popular XAI techniques—such as SHAP or Integrated Gradients—are justified by appealing to axioms like additivity or sensitivity. While these axioms offer important normative guidance, they do not ensure that all desirable explanations can be formally derived. Our result puts this on a rigorous foundation: there will always be edge cases where formal justification fails, even if the explanation is semantically valid.

We also note a limitation in epistemic assumptions. The original formulation of Axiom A7 claimed that all valid explanations are necessarily knowable to the user. This is too strong in many practical contexts—explanations may be semantically correct but inaccessible due to model complexity, opacity, or cognitive overload. To address this, we propose a refined epistemic axiom that only requires *possible* knowability (via modal logic). This provides a more realistic foundation for linking formal explanation validity to human interpretability.

10.3 Probabilistic and Causal Extensions

In Section 8, we extended our framework to capture probabilistic and causal semantics. These enhancements enable reasoning about uncertainty, interventions, and counterfactuals within the formal system. Importantly, we refined the notion of causal minimality to avoid incorrect sufficiency claims by ensuring minimality is defined via subset exclusion rather than individual feature relevance alone. These updates allow the system to represent more realistic explanation scenarios while preserving the underlying incompleteness result.

10.4 Proof Search and Automation

Section 9 introduced a backward-chaining proof search algorithm that explores the derivability of explanation statements using inference rules and axioms. While the algorithm cannot overcome the limits established by incompleteness, it serves as a basis for semi-automated explanation reasoning and verification. Formalizing unification and defining bounded proof strategies opens a path toward practical tooling for checking explainability claims.

10.5 Meta-Systems and External Reasoning

One possible path forward is the development of *meta-explanation systems*—reasoning frameworks that operate on the outputs or limitations of first-order explanation systems. These could incorporate human input, empirical heuristics, or higher-order logic to reason about the validity of explanations not provable within the base system $\mathcal{E}$. This is analogous to how mathematicians often reason informally about statements outside formal systems.

10.6 Relation to Human-Centric Interpretability

Humans frequently accept explanations that are plausible, incomplete, or not rigorously justified. The incompleteness of $\mathcal{E}$ reflects a similar reality: that full transparency may be logically unreachable. Instead of seeking a single unified explanation mechanism, future systems might embrace a pluralistic view—combining formal, empirical, and interactive modes of interpretability.

10.7 Ethical and Practical Implications

Recognizing the limits of explainability is critical in high-stakes domains. Over-reliance on formal systems may lead to a false sense of certainty. Our result reinforces the need for transparency not only in what models do, but also in what explanation frameworks can and cannot reveal.

10.8 Toward a Layered Architecture for Explainability

Taken together, our findings suggest that explanation in machine learning is inherently multi-layered. Formal logic provides structure and rigor, but must be complemented by causal insight, computational reasoning, and user-centered interaction. A complete account of explainability may thus require tools that integrate these perspectives into a modular, extensible framework—one that acknowledges the limitations of any single approach.

Acknowledgments and Disclosure of Funding

The author conducted this research independently but gratefully acknowledges the intellectual influence of foundational works in explainable AI and formal reasoning. In particular, the contributions of S. M. Lundberg and S.-I. Lee (SHAP), M. T. Ribeiro et al. (LIME), F. Doshi-Velez and B. Kim (interpretability formalism), and Judea Pearl (causal reasoning) provided important conceptual grounding. The formal perspective was further shaped by

the logic-based approaches of Halpern, Pearl, and Bringsjord et al. Any errors or interpretations remain the sole responsibility of the author.

References

Selmer Bringsjord, Naveen Sundar Govindarajulu, and John Taylor. Gödel, minds, and machines. *Journal of Applied Logics*, 8(1):43–80, 2021.

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

Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für Mathematik und Physik*, 38:173–198, 1931.

Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Penguin Books, 2019.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. Explainable ai: Interpreting, explaining and visualizing deep learning. In *Lecture Notes in Computer Science (LNCS)*, volume 11700, pages 1–438. Springer, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328, 2017.