Axiomatic Foundations of Counterfactual Explanations

Leila Amgoud
CNRS – IRIT
Toulouse, France
leila.amgoud@irit.fr

Martin Cooper
UPS – IRIT
Toulouse, France
martin.cooper@irit.fr

ABSTRACT

Explaining autonomous and intelligent systems is critical in order to improve trust in their decisions. *Counterfactuals* have emerged as one of the most compelling forms of explanation. They address "why not" questions by revealing how decisions could be altered. Despite the growing literature, most existing explainers focus on a single *type* of counterfactual and are restricted to *local* explanations, focusing on individual instances. There has been no systematic study of alternative counterfactual types, nor of *global* counterfactuals that shed light on a system's overall reasoning process.

This paper addresses the two gaps by introducing an axiomatic framework built on a set of desirable properties for counterfactual explainers. It proves impossibility theorems showing that no single explainer can satisfy certain axiom combinations simultaneously, and fully characterizes all compatible sets. Representation theorems then establish five one-to-one correspondences between specific subsets of axioms and the families of explainers that satisfy them. Each family gives rise to a distinct type of counterfactual explanation, uncovering five fundamentally different types of counterfactuals. Some of these correspond to local explanations, while others capture global explanations. Finally, the framework situates existing explainers within this taxonomy, formally characterizes their behavior, and analyzes the computational complexity of generating such explanations.

KEYWORDS

Explainable AI; Foundations of Explainability; Counterfactuals

ACM Reference Format:
Leila Amgoud and Martin Cooper. 2026. Axiomatic Foundations of Counterfactual Explanations. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/CKPK5561

1 INTRODUCTION

Explaining the decisions of autonomous and intelligent systems has attracted significant attention over the past decade (see [33, 34] for surveys). The overarching goal of this research is to enhance trust in the decisions made by such systems. Two primary types of explanations have been extensively studied. The first type addresses "why" questions by identifying the key factors that led to a specific decision (e.g., [10, 13, 32]). For example: "Alice is not eligible for a loan because *her annual income is* \$20K." The second type—*counterfactuals*—answers "why not" questions by describing how the decision could be changed. For instance: "If Alice's annual income had been \$35K, she would have been eligible for a loan."

Byrne argues in [6, 7] that individuals affected by decisions often prefer explanations that show how they can achieve better outcomes. This insight has motivated the development of numerous counterfactual explainers (see [15, 22, 24] for recent surveys), applicable to mono-agent and multi-agent systems (e.g., [9, 36]).

The vast majority of existing work has focused on *local counterfactuals*—explanations tailored to individual input instances—and has typically considered only a **single type of counterfactual**. A counterfactual for a given input is defined as the "closest" alternative input that yields a different outcome. The various explainers differ in how they formalize this notion of closeness. To guide its definition and the selection of "reasonable" counterfactuals, several properties have been proposed in the literature (e.g., [5, 16, 40, 46, 47]). For instance, *minimality* requires that a counterfactual involves the smallest possible change to the original input, while *diversity* promotes generating distinct counterfactuals for the same input. Existing explainers are typically evaluated based on the extent to which they satisfy these properties and on the strategies they employ to generate counterfactuals.

Despite a rich body of work on counterfactual explanations, there has been no systematic study of alternative types of local counterfactuals. Moreover, global counterfactuals—which aim to capture and explain a model's overall decision-making behavior—have received far less attention; to date, the only notable contribution is that of [2]. As a result, the theoretical trade-offs between local and global counterfactuals remain poorly understood.

This paper addresses these gaps by proposing a formal framework for systematically defining and comparing different types of local and global counterfactuals. Our approach is axiomatic—an established method in modern theoretical science that builds systems upon clearly stated axioms (basic properties) from which representation theorems can be derived. Applying this approach to counterfactuals offers several benefits. First, it brings **clarity** and **rigor** by precisely specifying desirable properties, fostering a shared understanding of what constitutes a sound counterfactual. Second, it enables a **systematic comparison** of counterfactual types based on the axioms they satisfy, making their differences, strengths, and limitations explicit. This supports a principled and objective evaluation that goes beyond empirical performance alone. Finally, the framework is **model-agnostic** and **general**, ensuring that the results are broadly applicable across diverse settings.

The contributions of this paper are sixfold and complement the existing literature on counterfactual explanations:

(1) It introduces an *axiomatic framework* grounded on a set of nine axioms.
(2) It provides *impossibility theorems*, showing that no explainer can satisfy certain combinations of axioms simultaneously.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/CKPK5561

(3) It fully characterizes all compatible sets of axioms.
(4) It presents *representation theorems* that define one-to-one correspondences between specific subsets of axioms and the families of counterfactual explainers that satisfy them. These characterizations are valuable because they precisely and **exhaustively** describe **all** explainers that meet a given set of axioms. Key findings include:
   - The nine axioms distinguish five distinct types of counterfactuals, each generated by a different family of explainers.
   - There are two fundamental forms of counterfactual explanation: *necessary* and *sufficient* reasons.
   - For each form (necessary, sufficient), one type corresponds to *global* explanations, while the remaining types focus on *local* explanations.
   - Among local sufficient reasons, there are two subtypes: *sceptical* and *credulous*.
(5) It formally characterizes existing counterfactual explainers, showing that nearly all fall within the family generating local credulous sufficient reasons. The global explainers proposed in [2] correspond to the family of global necessary reasons. Thus, the framework uncovers three previously unrecognized types of counterfactuals.
(6) It analyzes the computational complexity of generating explanations.

The paper is organized as follows. Section 2 presents the necessary background, Section 3 defines axioms, Sections 4 and 5 are devoted to representation theorems. Section 6 compares the types, Section 7 discusses related work, and Section 8 provides complexity results.

2 BACKGROUND

Throughout the paper, we consider a *classification theory* as a tuple $\mathbb{T} = \langle \mathtt{F}, \mathtt{dom}, \mathtt{C} \rangle$ comprising a finite set $\mathtt{F}$ of *features*, a function $\mathtt{dom}$ which returns the *domain* of every feature $f \in \mathtt{F}$ such that $\mathtt{dom}(f)$ is finite and $|\mathtt{dom}(f)| > 1$, and a finite set $\mathtt{C}$ of *classes* with $|\mathtt{C}| \geq 2$. Let $\mathtt{Lit}(\mathbb{T})$ be the set of all literals in $\mathbb{T}$, where a *literal* is a pair (f, v) where $f \in \mathtt{F}$ and $v \in \mathtt{dom}(f)$. A *partial assignment* is any set of literals with each feature in $\mathtt{F}$ occurring at most once; it is called an *instance* when every feature appears once. We denote by $\mathbb{E}(\mathbb{T})$ the set of all possible partial assignments and by $\mathbb{F}(\mathbb{T})$ the *feature space*, i.e., the set of all instances, of theory $\mathbb{T}$. For a set $E \in \mathbb{E}(\mathbb{T})$, $\mathbb{E}(\mathbb{T}) \oslash E$ denotes the set of all partial assignments that do not contain any literal from E, i.e., for any $X \in \mathbb{E}(\mathbb{T}) \oslash E$, $X \in \mathbb{E}(\mathbb{T})$ and $E \cap X = \emptyset$. For $E \in \mathbb{E}(\mathbb{T})$ and $x \in \mathbb{F}(\mathbb{T})$, we denote by $x_{\downarrow E}$ the partial assignment such that $x_{\downarrow E} = E \cup \{(f, v) \in x \mid \nexists v' \in \mathtt{dom}(f) \text{ s.t. } v' \neq v \text{ and } (f, v') \in E\}$. The idea is to replace the values of common features by those in E. Consider a theory consisting of two binary features, the instance $x = ((f_1, 0), (f_2, 0))$ and $E = \{(f_2, 1)\}$. Then, $x_{\downarrow E} = ((f_1, 0), (f_2, 1))$.

Property 1. *Let $\mathbb{T}$ be classification theory. $\forall x \in \mathbb{F}(\mathbb{T}), \forall E \in \mathbb{E}(\mathbb{T})$, $x_{\downarrow E} \in \mathbb{F}(\mathbb{T})$.*

For any $x \in \mathbb{F}(\mathbb{T})$ and any $E \in \mathbb{E}(\mathbb{T})$, $x \ominus E = \{y \in \mathbb{F}(\mathbb{T}) \mid x \setminus y = E\}$. Consider a theory consisting of two features, where $\mathtt{dom}(f_1) = \{0, 1\}$ and $\mathtt{dom}(f_2) = \{0, 1, 2\}$, the instance $x = ((f_1, 0), (f_2, 0))$ and $E = \{(f_2, 0)\}$. Then, $x \ominus E = \{y, z\}$ where $y = ((f_1, 0), (f_2, 1))$ and $z = ((f_1, 0), (f_2, 2))$.

A *classifier* on a theory $\mathbb{T}$ is a **surjective** function $\kappa : \mathbb{F}(\mathbb{T}) \to \mathtt{C}$ mapping every instance to a class such that every class is assigned to at least one instance.

Among literals, we distinguish those that are *core*, or mandatory, to a given class under a given classifier.

Definition 1. *Let $\mathbb{T} = \langle \mathtt{F}, \mathtt{dom}, \mathtt{C} \rangle$ be a theory, $c \in \mathtt{C}$, κ a classifier on $\mathbb{T}$, $l \in \mathtt{Lit}(\mathbb{T})$. l is* core *to c iff $\forall x \in \mathbb{F}(\mathbb{T})$ s.t. $\kappa(x) = c$, $l \in x$. $\mathtt{Core}(c)$ is the set of core literals of c.*

The set $\mathtt{Core}(c)$, for $c \in \mathtt{C}$, may be empty as shown below.

Example 1. *Let κ be a classifier that predicts where to take a friend. It is a function of the temperature t and the friend's favourite activity a, with $\mathtt{dom}(t) = \{hot, mild, freezing\}$, $\mathtt{dom}(a) = \{climbing, reading, skiing\}$, and $\mathtt{C} = \{beach, mountain, cinema\}$. κ predicts the beach on hot days, the mountain if it is mild and the friend likes climbing or it is freezing cold and she likes skiing, otherwise the cinema. Below are all instances and their decisions.*

$\mathbb{F}(\mathbb{T})$	t	a	$\kappa(x_i)$
x_1	*hot*	*climbing*	*beach*
x_2	*mild*	*climbing*	*mountain*
x_3	*freezing*	*reading*	*cinema*
x_4	*freezing*	*skiing*	*mountain*
x_5	*freezing*	*climbing*	*cinema*
x_6	*hot*	*reading*	*beach*
x_7	*hot*	*skiing*	*beach*
x_8	*mild*	*reading*	*cinema*
x_9	*mild*	*skiing*	*cinema*

The following lists the set of core literals corresponding to each class.

- $\mathtt{Core}(beach) = \{(t, hot)\}$,
- $\mathtt{Core}(mountain) = \emptyset$,
- $\mathtt{Core}(cinema) = \emptyset$.

Let X be a set of objects. A *preordering* on X is a binary relation $\succeq$ on X that is *reflexive* and *transitive*. It is *total* iff for all $x, y \in X$, $x \succeq y$ or $y \succeq x$. The notation $x \succeq y$ stands for x is at least as preferred as y; $x \succ y$ is a shortcut for $x \succeq y$ and $y \not\succeq x$. Let $\max(X, \succeq) = \{E \in X \mid \nexists E' \in X \text{ such that } E' \succ E\}$ be the set of most preferred elements of X. Finally, a *weighting* on the set X is a function $\sigma : X \to [0, +\infty)$.

3 AXIOMATIC FRAMEWORK

We introduce an axiomatic framework that defines counterfactual explainers–functions generating counterfactual explanations. Before presenting the axioms, we begin by introducing the notion of a *query*, which represents the question posed to an explainer.

Definition 2. *A* query *is a tuple $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ such that $\mathbb{T}$ is a classification theory, κ a classifier on $\mathbb{T}$ and $x \in \mathbb{F}(\mathbb{T})$.*

A *counterfactual explainer*—or explainer for short—is a function that maps each query to a set of partial assignments, each representing a *counterfactual*. As we will see in the following sections, there is no single definition of a counterfactual. Nevertheless, it is widely acknowledged in the literature that all such definitions aim to provide insights into how a prediction made by a classifier (or an AI system) can be altered.

General	*Success*	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\mathbf{L}(\mathbf{Q}) \neq \emptyset$.
	Non-Triviality	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\emptyset \notin \mathbf{L}(\mathbf{Q})$.
	Equivalence	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\forall \mathbf{Q}' = \langle \mathbb{T}, \kappa, x' \rangle$, if $\kappa(x) = \kappa(x')$, then $\mathbf{L}(\mathbf{Q}) = \mathbf{L}(\mathbf{Q}')$.
Nec. Reasons	*Feasibility*	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\forall E \in \mathbf{L}(\mathbf{Q})$, $E \subseteq x$.
	Coreness	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\forall E \in \mathbf{L}(\mathbf{Q})$, $E \subseteq \texttt{Core}(\kappa(x))$.
	Sceptical Validity	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\forall E \in \mathbf{L}(\mathbf{Q})$, $\forall y \in x \ominus E$, $\kappa(y) \neq \kappa(x)$.
Suff. Reasons	*Novelty*	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\forall E \in \mathbf{L}(\mathbf{Q})$, $x \cap E = \emptyset$.
	Strong Validity	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\forall E \in \mathbf{L}(\mathbf{Q})$, $\nexists y \in \mathbb{F}(\mathbb{T})$ s.t. $E \subseteq y$ and $\kappa(y) = \kappa(x)$.
	Weak Validity	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$, $\forall E \in \mathbf{L}(\mathbf{Q})$, $\kappa(x_{\downarrow E}) \neq \kappa(x)$.

Table 1: Axioms for explainer L.

Definition 3. *A* counterfactual explainer *is a function* $\mathbf{L}$ *mapping every query* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *into a subset of* $\mathbb{E}(\mathbb{T})$. *Every* $E \in \mathbf{L}(\mathbf{Q})$ *is a* counterfactual.

Notation: Let $\mathbf{L}, \mathbf{L}'$ be two explainers. We write $\mathbf{L} \sqsubseteq \mathbf{L}'$ to denote that for any query $\mathbf{Q}$, $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}'(\mathbf{Q})$.

Table 1 introduces axioms—formal properties of explainers—three of which, Success, Non-Triviality, and Coreness, are borrowed from [2], while the remaining ones are novel. The axioms are grouped into three categories: the first contains general properties that any explainer might satisfy, while the other two define distinct forms of counterfactuals as will be shown in next sections.

Success guarantees at least one explanation per query; *Non-Triviality* rejects empty explanations since uninformative; and *Equivalence* ensures that instances with the same decisions get the same explanations. This axiom is suitable for explainers that produce global explanations—i.e., those that concern classes rather than input instances. *Feasibility* states that an explanation must be part of the instance whose decision is being explained. As we will see later, for some type of counterfactuals, this axiom helps identify the features whose values need to be changed in order to alter the instance's prediction. *Coreness* goes further by ensuring that an explanation contains only core literals for the decision. The logic is that missing a core literal certainly leads to another class. *Sceptical validity* ensures that an explanation E surely leads to a change in a decision. Indeed, every instance that differs from the one being explained only in E has a different class. *Novelty* ensures that an explanation does not share a literal with the original instance. The idea is that in order to inform a user how to change an output, the explanation must involve modifications to the input. This axiom concerns some types of counterfactuals. *Strong validity* is similar in spirit to Sceptical validity as it ensures that an explanation E must lead to a change in a decision. Indeed, any instance that contains E has a different prediction. *Weak Validity* is less demanding, since it only requires that the instance obtained by replacing the new literals in the original instance is of a different class.

The axioms are not all independent, as demonstrated below. Nevertheless, each plays a crucial role in defining distinct types of counterfactual explanations.

Proposition 1. *The following implications hold.*

- *Coreness* $\Rightarrow$ *Feasibility.*
- *Coreness and Non-Triviality* $\Rightarrow$ *Sceptical Validity.*
- *Sceptical Validity* $\Rightarrow$ *Non-Triviality.*
- *Weak Validity* $\Rightarrow$ *Non-Triviality.*
- *Strong Validity* $\Rightarrow$ *Weak Validity.*
- *Novelty and Non-Triviality* $\Rightarrow$ *Sceptical Validity.*

We present an impossibility theorem showing the *incompatibility* of certain axioms—no explainer can satisfy all of them simultaneously. As we will show later, each compatible subset gives rise to a distinct form of counterfactual.

Theorem 1. *The axioms of every set* $(\mathtt{I}_i)$ *are incompatible.*

- $(\mathtt{I}_1)$ *Success, Non-Triviality and Coreness.*
- $(\mathtt{I}_2)$ *Success, Feasibility and Sceptical Validity.*
- $(\mathtt{I}_3)$ *Success, Novelty and Strong Validity.*
- $(\mathtt{I}_4)$ *Success, Non-Triviality, Feasibility and Novelty.*
- $(\mathtt{I}_5)$ *Success, Feasibility and Weak Validity.*
- $(\mathtt{I}_6)$ *Success, Non-Triviality,Equivalence and Feasibility.*
- $(\mathtt{I}_7)$ *Success, Non-Triviality, Equivalence and Novelty.*

All combinations of axioms not excluded by the two preceding results are compatible; in other words, each such combination is satisfied by at least one explainer.

Theorem 2. *All combinations of axioms not disallowed by Proposition 1 and Theorem 1 are satisfied by some explainer.*

Two fundamental questions emerge concerning these axioms:

(1) Can we precisely and exhaustively characterize the family of **all** explainers that satisfy a given subset of the axioms?
(2) Do the axioms lead to the same type of counterfactuals?

In the next two sections, we present representation theorems that answer the first question affirmatively and reveal the existence of **five distinct types of counterfactuals**. These theorems show that counterfactual explanations fall into two fundamental forms—*necessary* and *sufficient*—each encompassing several distinct types. For clarity, we dedicate a separate section to each form.

4 NECESSARY REASONS

There first form of counterfactuals can be expressed as follows: "*If it were not the case that E, the decision would have been different*". For example, in instance x_1 of Example 1 if the temperature had not been hot, the decision would have been different, as can be verified from the table. Such explanations indicate the **part of an instance** that should be changed to avoid its decision. We call them *necessary reasons* since they are mandatory to guarantee the current decision.

In this section, we study their *global* and *local* versions. The former explain a class while latter target an instance.

4.1 Global Necessary Reasons

Global necessary reasons explain the behavior of a classifier, that is, how it assigns classes. For a given class, they identify feature–value combinations that **distinguish it from all other classes**. In their absence, the classifier assigns a different class.

Definition 4. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query. A* global necessary reason *(GNR) for* $\kappa(x)$ *is a set* $E \in \mathbb{E}(\mathbb{T})$ *such that:*

$$E \neq \emptyset \quad \textit{and} \quad \forall y \in \mathbb{F}(\mathbb{T}),\ \textit{if}\ E \not\subseteq y,\ \textit{then}\ \kappa(y) \neq \kappa(x).$$

gNec *denotes the explainer that returns all GNRs.*

Ex. 1 (Cont) Let $\mathbf{Q}_i$ be the query on instance x_i, $i = 1, 2, 3$.

- $\mathtt{gNec}(\mathbf{Q}_1) = \{\{(t, \text{hot})\}\}$ $\mathtt{gNec}(\mathbf{Q}_2) = \mathtt{gNec}(\mathbf{Q}_3) = \emptyset$.

A hot temperature is characteristic of the *beach* option, as it sets it apart from the mountain and cinema options. Thus, to change the decision for x_1, the value of t must be altered to either mild or freezing. In contrast, the other two queries have no GNR, since every feature–value combination appears in at least two instances that are assigned different classes. For example, modifying the value of a for x_3 does not necessarily lead to a class change, as $\kappa(x_5) = \kappa(x_3)$.

We provide a representation theorem that characterizes the **entire family** of explainers generating GNRs. These are the **only ones** that satisfy both Coreness and Non-triviality.

Theorem 3. *An explainer* $\mathbf{L}$ *satisfies Coreness and Non-triviality* ***iff*** $\mathbf{L} \sqsubseteq \mathtt{gNec}$.

Table 2 summarizes all the properties that are satisfied/violated by gNec and any explainer $\mathbf{L}$ such that $\mathbf{L} \sqsubseteq \mathtt{gNec}$.

Theorem 4. *The properties of Table 2 hold.*

These explainers violate Success since Theorem 1 shows that Coreness, Success and Non-Triviality are incompatible. Since gNec generates global explanations, it satisfies Equivalence: its explanations are common to all instances labelled by the same class. For example, $\mathtt{gNec}(\mathbf{Q}_1) = \mathtt{gNec}(\mathbf{Q}_6) = \mathtt{gNec}(\mathbf{Q}_7) = \{\{(t, \text{hot})\}\}$, where $\mathbf{Q}_i$ concerns the instance x_i.

To sum up, the characterization result shows that the family of explainers satisfying the Coreness and Non-triviality axioms generates global necessary reasons for a class. This type of counterfactual highlights the factors that distinguish a given class from **all others**.

4.2 Local Necessary Reasons

We now turn to the problem of explaining the decision made for a specific instance. A natural approach is to consider the global reasons associated with the instance's class. However, this approach is overly rigid, as it requires exploring the entire feature space, which drastically reduces the chances of finding a counterfactual (e.g., x_2, x_3 in Example 1). In what follows, we introduce **sceptical necessary reasons** (SNR), an alternative approach that is less demanding. SNRs are subsets of an instance of interest x that distinguish it from **any** other instance y for which the decision would be different. Thus, their absence guarantees a different decision for x. Unlike the global approach, the local approach restricts the search space to instances that differ from x only in candidate reasons.

Definition 5. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query. A* sceptical necessary reason *(SNR) for* $\kappa(x)$ *is a set* $E \in \mathbb{E}(\mathbb{T})$ *such that:*

$$E \subseteq x \qquad \textit{and} \qquad \forall y \in x \ominus E,\ \kappa(y) \neq \kappa(x).$$

sNec *denotes the explainer that generates all SNRs.*

Ex. 1 (Cont) Let $\mathbf{Q}_i = \langle \mathbb{T}, \kappa, x_i \rangle$.

- $\mathtt{sNec}(\mathbf{Q}_1) = \{\{(t, \text{hot})\}, x_1\}$.
- $\mathtt{sNec}(\mathbf{Q}_2) = \{\{(t, \text{mild})\}, \{(a, \text{climbing})\}\}$ $\mathtt{sNec}(\mathbf{Q}_3) = \emptyset$.

Note that $x_2 \ominus \{(t, \text{mild})\} = \{x_1, x_5\}$, $\kappa(x_2) \neq \kappa(x_1)$, $\kappa(x_2) \neq \kappa(y)$. Hence, the SNR $\{(t, \text{mild})\}$ distinguishes x_2 from all instances that differ from it **only** in temperature.

Note that global necessary reasons are sceptical ones. Consequently, if there are core literals, their non-empty subsets are SNRs. But, an explainer may return additional reasons whose literals are not core to the instance class (e.g., the case of $\mathbf{Q}_2$).

Proposition 2. *It holds that* $\mathtt{gNec} \sqsubseteq \mathtt{sNec}$.

We present a representation theorem showing an equivalence between the two axioms (Feasibility, Sceptical Validity) and the family of explainers that generate SNRs.

Theorem 5. *An explainer* $\mathbf{L}$ *satisfies Feasibility and Sceptical Validity* ***iff*** $\mathbf{L} \sqsubseteq \mathtt{sNec}$.

Table 2 summarizes the axioms that are satisfied/violated by sNec and, more generally, by any explainer $\mathbf{L} \sqsubseteq \mathtt{sNec}$. Note that they satisfy non-Triviality, hence they are **informative**. However, they violate Success and so might not guarantee explanations for queries (eg. $\mathbf{Q}_3$). They also violate Coreness but satisfy its weaker version, Sceptical Validity. The latter guarantees a change in decision no matter what the new values of the features are, making SNRs **valid** in the sense of the property defined in [15].

To sum up, Theorem 5 shows that the family of explainers satisfying Feasibility and Sceptical Validity generates local necessary reasons for an instance. Such reasons identify the factors that distinguish the instance from all others receiving a different decision.

5 SUFFICIENT REASONS

The second form of counterfactuals can be expressed as follows: "*If it were the case that E, the decision would have been different.* " Unlike necessary reasons, which identify present characteristics whose absence would change the outcome, this second type focuses on absent characteristics whose presence would alter the outcome. For instance, "If Alice's annual income had been $\$35K$ (instead of the actual amount), she wouldn't have been denied a loan". We introduce the concept of *sufficient reasons*—feature values that directly lead to a different decision—rather than identifying which parts of an instance must be modified.

5.1 Global Sufficient Reasons

We begin by analyzing the global behavior of a classifier and, consequently, of its classes. As previously noted, global necessary reasons capture what distinguishes one class from the others; however, such reasons may not always exist, as illustrated by instances x_2 and x_3 in Example 1. To address this, we introduce the notion of *global sufficient reasons* for *avoiding* a class—factors whose presence leads the classifier to assign a different class.

	Necessary Reasons				Sufficient Reasons					
Axioms	$\mathbf{L} \sqsubseteq$ gNec	gNec	$\mathbf{L} \sqsubseteq$ sNec	sNec	$\mathbf{L} \sqsubseteq$ gSuf	gSuf	$\mathbf{L} \sqsubseteq$ sSuf	sSuf	$\mathbf{L} \sqsubseteq$ cSuf	cSuf
Success	×	×	×	×	–	✓	×	×	–	✓
Non-Triviality	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Equivalence	–	✓	–	×	–	✓	–	×	–	×
Feasibility	✓	✓	✓	✓	–	×	–	×	–	×
Coreness	✓	✓	–	×	–	×	×	×	×	×
Sceptical Validity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Novelty	–	×	–	×	–	×	✓	✓	✓	✓
Strong Validity	–	×	–	×	✓	✓	✓	✓	–	×
Weak Validity	–	×	–	×	✓	✓	✓	✓	✓	✓

Table 2: Formal Comparison of Explainers. "–" means the axiom is satisfied by only some instances of the family but not all.

Definition 6. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query. A* global sufficient reason *(GSR) against* $\kappa(x)$ *is a set* $E \in \mathbb{E}(\mathbb{T})$ *such that:*

$$\forall y \in \mathbb{F}(\mathbb{T}) \text{ with } E \subseteq y,\ \kappa(y) \neq \kappa(x).$$

Let gSuf *be the explainer that generates all GSRs.*

Ex. 1 (Cont) Let $\mathbf{Q}_i = \langle \mathbb{T}, \kappa, x_i \rangle$, $i = 1, 2, 3$.

- $\texttt{gSuf}(\mathbf{Q}_1) = \{\{(t, \text{mild})\}, \{(t, \text{freezing})\}\} \cup \{x_2, x_3, x_4, x_5, x_8, x_9\}$.
- $\texttt{gSuf}(\mathbf{Q}_2) = \{\{(t, \text{hot})\}, \{(a, \text{reading})\}\} \cup \{x_1, x_3, x_5, x_6, x_7, x_8, x_9\}$.
- $\texttt{gSuf}(\mathbf{Q}_3) = \{\{(t, \text{hot})\}\} \cup \{x_1, x_2, x_4, x_6, x_7\}$.

Note that $\mathbf{Q}_1$ has 8 GSRs; the first one reads: *If the temperature had been mild, the decision would have been different from beach.*

The following representation theorem shows that only explainers satisfying Strong Validity produce (GSRs).

Theorem 6. *An explainer* $\mathbf{L}$ *satisfies Strong Validity* ***iff*** $\mathbf{L} \sqsubseteq$ gSuf.

Table 2 summarizes the behavior of gSuf and any explainer $\mathbf{L} \sqsubseteq$ gSuf with respect to all axioms. It confirms that the two forms of counterfactuals—necessary and sufficient—convey distinct types of information, as they satisfy different subsets of axioms. Unlike gNec, gSuf guarantees at least one non-empty explanation for any query but violates Feasibility.

To sum up, Theorem 6 shows that the axiom of Strong Validity gives rise to a new type of global counterfactuals—those that highlight feature–value combinations whose presence guarantees the avoidance of a given class.

5.2 Local Sufficient Reasons

Let us now explain decisions of individual instances. We distinguish two types of explanations: *sceptical* and *credulous*.

5.2.1 Sceptical Sufficient Reasons (SSRs). They highlight features that are characteristic of instances assigned to classes different from that of the instance under consideration. As such, they act as distinguishing features of other classes and must be absent from the instance of interest. Consequently, SSRs consist exclusively of **new feature values** not already present in the instance.

Definition 7. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query. A* sceptical sufficient reason *(SSR) against* $\kappa(x)$ *is a set* $E \in \mathbb{E}(\mathbb{T})$ *s.t.:*

$$E \cap x = \emptyset \quad \text{and} \quad \forall y \in \mathbb{F}(\mathbb{T}) \text{ s.t. } E \subseteq y,\ \kappa(y) \neq \kappa(x).$$

sSuf *denotes the explainer that generates* all *SSRs.*

Ex. 1 (Cont) Let $\mathbf{Q}_i = \langle \mathbb{T}, \kappa, x_i \rangle$, $i = 1, 2, 3$.

- $\texttt{sSuf}(\mathbf{Q}_1) = \{\{(t, \text{mild})\}, \{(t, \text{freezing})\}\} \cup \{x_3, x_4, x_8, x_9\}$.
- $\texttt{sSuf}(\mathbf{Q}_2) = \{\{(t, \text{hot})\}, \{(a, \text{reading})\}\} \cup \{x_3, x_6, x_7\}$.
- $\texttt{sSuf}(\mathbf{Q}_3) = \{\{(t, \text{hot})\}\} \cup \{x_1, x_2, x_7\}$.

Consider, for instance, the reason $\{(t, \text{mild})\}$. This feature–value pair characterizes instances labeled either mountain or cinema. There is no instance labeled beach with a mild temperature.

The next representation theorem establishes a one-to-one correspondence between the two axioms Novelty and Strong Validity, and the family of explainers that generate SSRs.

Theorem 7. *An explainer* $\mathbf{L}$ *satisfies Novelty and Strong Validity* ***iff*** $\mathbf{L} \sqsubseteq$ sSuf.

Table 2 shows that this fourth type of counterfactuals is different from local necessary reasons. However, both types may not exist.

To sum up, the above characterization shows that Strong Validity and Novelty define a fourth type of counterfactual, based on absent feature–value combinations that always yield a different decision.

5.2.2 Credulous Sufficient Reasons (CSRs). Being the least restrictive type of reason, CSRs are feature–value combinations that distinguish an instance from **another instance** classified differently. Unlike SSRs, the presence of a CSR does not necessarily guarantee a change of decision. In the running example, *skiing* is a CSR for instance x_3, since changing the value of a to *skiing* yields x_4, which receives a different decision. However, *skiing* is not an SSR, as it also appears in x_9, whose class is the same as x_3.

Definition 8. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query. A* credulous sufficient reason against $\kappa(x)$ *is* $y \setminus x$ *such that* $y \in \mathbb{F}(\mathbb{T})$ *and* $\kappa(y) \neq \kappa(x)$. cSuf *is the explainer that generates all CSRs.*

Ex. 1 (Cont) Let $\mathbf{Q}_i = \langle \mathbb{T}, \kappa, x_i \rangle$, $i = 1, 2, 3$.

- $\texttt{cSuf}(\mathbf{Q}_1) = \texttt{sSuf}(\mathbf{Q}_1)$.
- $\texttt{cSuf}(\mathbf{Q}_2) = \{\{(t, \text{hot})\}, \{(t, \text{freezing})\}, \{(a, \text{reading})\}, \{(a, \text{skiing})\}\} \cup \{y \in \mathbb{F}(\mathbb{T}) \mid y \cap x_2 = \emptyset \text{ and } \kappa(y) \neq \text{mountain}\}$.
- $\texttt{cSuf}(\mathbf{Q}_3) = \{\{(t, \text{hot})\}, \{(a, \text{skiing})\}\} \cup \{y \in \mathbb{F}(\mathbb{T}) \mid y \cap x_3 = \emptyset \text{ and } \kappa(y) \neq \text{cinema}\}$.

Note that $\{(a, \text{skiing})\}$ is not a SSR of $\mathbf{Q}_3$ since skiing is not representative of mountain or beach. Indeed, κ predicts cinema (class of x_3) for the instance $x_9 = ((t, \text{mild}), (a, \text{skiing}))$.

In what follows, we present a representation theorem showing that the axioms Novelty and Weak Validity are satisfied exclusively by explainers that return CSRs.

Theorem 8. *An explainer* $\mathbf{L}$ *satisfies Novelty and Weak Validity* ***iff*** $\mathbf{L} \sqsubseteq \texttt{cSuf}$.

We now introduce the concept of a **faithful ranking**—a preorder over the set of partial assignments that strictly favors those introducing new feature values to an instance, where such revisions lead to a different decision.

Definition 9. *A* faithful ranking *is a function* $\mathbf{S}$ *that maps every query* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *to a preorder* $\succeq_{\mathbf{Q}}$ *over* $\mathbb{E}(\mathbb{T})$ *s.t* $\forall E, E' \in \mathbb{E}(\mathbb{T})$, *if:*

- $E \cap x = \emptyset$ *and* $\kappa(x_{\downarrow E}) \neq \kappa(x)$ *and*
- $E' \cap x \neq \emptyset$ *or* $\kappa(x_{\downarrow E'}) = \kappa(x)$,

then $E \succ_{\mathbf{Q}} E'$.

In what follows, we present a second representation theorem that connects the axioms Success, Novelty and Weak Validity with credulous sufficient reasons. It establishes an equivalence between these axioms and an explanation strategy grounded in preorders over partial assignments. Specifically, explainers satisfying these axioms impose a ranking over all partial assignments—from least to most preferred—and select the top-ranked ones as explanations.

Theorem 9. *An explainer* $\mathbf{L}$ *satisfies Success, Novelty and Weak Validity* ***iff*** *there exists a faithful ranking* $\mathbf{S}$ *mapping every query* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *to a preorder* $\succeq_{\mathbf{Q}}$ *on* $\mathbb{E}(\mathbb{T})$ *such that:*

$$\mathbf{L}(\mathbf{Q}) = \max(\mathbb{E}(\mathbb{T}), \succeq_{\mathbf{Q}}).$$

The above result characterizes the sub-family of explainers $\mathbf{L} \sqsubseteq \texttt{cSuf}$, identified in Theorem 8, that guarantees the existence of at least one CSR. In a later section, we will show that this characterization, expressed in terms of faithful rankings, facilitates the comparison of existing explainers that generate CSRs.

The next result relates the three types of sufficient reasons (GSR, SSR, CSR) and shows that every local explanation is a subset of a global explanation.

Proposition 3. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query.*

- $\texttt{sSuf} \sqsubseteq \texttt{gSuf}$.
- $\texttt{sSuf} \sqsubseteq \texttt{cSuf}$.
- *For any* $E \in \texttt{cSuf}(\mathbf{Q})$, $\exists E' \in \texttt{gSuf}(\mathbf{Q})$ *s.t.* $E = E' \setminus x$.
- *For any* $E \in \texttt{gSuf}(\mathbf{Q})$, $E \setminus x \in \texttt{cSuf}(\mathbf{Q})$.

The properties summarized in Table 2 distinguish this fifth type of counterfactual from the other four. Unlike SSR, CRS are guaranteed to exist. Consequently, well-defined explainers can provide an outcome for every query.

6 HUMAN-FOCUSED ASSESSMENT OF TYPES

Cognitive science research shows that people generally prefer explanations that are short and concise [7, 8, 45]. We examine how the five types of counterfactuals align with these needs, using the following criteria for comparison.

Existence: Whether counterfactuals are guaranteed to exist.

Compactness: The ability to convey multiple pieces of information in a concise way.

Size: The number of literals in a counterfactual.

Number: The number of explanations provided for a query.

Power of discrimination: The ability of a counterfactual to discriminate between instances or classes.

Table 2 shows that necessary reasons and sceptical sufficient reasons may not always exist, whereas (global, credulous) sufficient reasons guarantee at least one explanation for any query. While the potential non-existence of reasons could be viewed as a drawback, [15] highlights that features in credulous sufficient reasons may not be *actionable* (e.g., age cannot be changed). Therefore, an explainer $\mathbf{L} \sqsubseteq \texttt{cSuf}$ generating CSRs composed of features to be mutated may face cases where counterfactuals do not exist.

When they do exist, necessary reasons (whether local or global) represent feature–value combinations whose absence alone causes a change in class, regardless of the new values assigned to these features. Hence, it is unnecessary to list all possible changes that could lead to a different decision. This allows necessary reasons to convey multiple pieces of information—the full range of changes—in a **compact** or concise way, making them particularly appealing for people. In contrast, sufficient reasons are not compact, as they enumerate all possible changes explicitly.

Because of their compactness, necessary reasons are typically **fewer in number** than sufficient reasons. In the running example, $\texttt{gNec}$ returns a single explanation for query $\mathbf{Q}_1$, whereas $\texttt{gSuf}$ yields eight GSRs. In fact, each alternative value of a feature occurring in a GNR gives rise to distinct GSRs. Global necessary reasons are also **shorter** than global sufficient reasons, since each core literal provides a minimal explanation.

Global counterfactuals are **discriminating**, as they capture feature–value pairs that distinguish one class from all the others. Likewise, sceptical (necessary, sufficient) reasons identify combinations that set an instance apart from others. CSRs, however, are less discriminating, as they may include pairs also found in instances with the same outcome. For instance, (t,freezing) is a CSR for x_2 but also belongs to x_4, which has the same decision as x_2.

In conclusion, necessary reasons, whenever they exist, seem more compatible with human preferences than sufficient reasons.

7 RELATED WORK

Numerous studies have investigated counterfactual explanations (see [15] for a recent survey). Some contributions propose new explainers, while others focus on identifying desirable properties that counterfactuals and the explainers that generate them should satisfy. In the following, we review both lines of work.

7.1 Counterfactuals Types and Explainers

Despite the rich body of work on counterfactual explanations, the vast majority has focused on **local** counterfactuals (e.g., [3, 11, 12, 14, 25, 31, 35, 37, 41, 44]). Furthermore, given an input x, these works generate the "closest" instance y that gets a different decision. The main distinction between the proposals lies in how closeness is defined. Some emphasize **sparsity**, seeking to minimize the set of features to be changed, while others prioritize **proximity**, allowing broader changes as long as the overall shift remains small according

to a specified **distance metric** like *Manhattan distance* (L_1 norm) or *Gower distance* (see [44] for more examples). We show that all these explainers generate **credulous sufficient reasons**, thus are specific instances within the family of CSRs. In other words, each such explainer, say $\mathbf{L}$, satisfies the condition: $\mathbf{L} \sqsubseteq \mathtt{cSuf}$. Furthermore, we characterize the faithful ranking defined by each explainer.

There are two explainers that promote sparsity. The first generates contrastive explanations [11]—**minimal sets of features** that must be altered to get a different decision. For the sake of comparison, we present its corresponding explainer whose explanations are sets of literals (not features).

Definition 10. *We define* $\mathbf{L_{wf}}$ *as the explainer that, for any query* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *and* $E \in \mathbb{E}(\mathbb{T})$, $E \in \mathbf{L_{wf}}(\mathbf{Q})$ *iff:*

- $\kappa(x_{\downarrow E}) \neq \kappa(x)$,
- $\nexists E' \in \mathbb{E}(\mathbb{T})$ *s.t.* $\kappa(x_{\downarrow E'}) \neq \kappa(x)$, $\mathtt{Feat}(E') \subset \mathtt{Feat}(E)$ *(with* $\mathtt{Feat}(E)$ *the set of* ***features*** *covered in* E*).*

Ex. 1 (Cont) Let $\mathbf{Q}_i = \langle \mathbb{T}, \kappa, x_i \rangle$, $i = 1, 2, 3$.

- $\mathbf{L_{wf}}(\mathbf{Q}_1) = \{\{(t, \text{mild})\}, \{(t, \text{freezing})\}\}$,
- $\mathbf{L_{wf}}(\mathbf{Q}_2) = \{\{(t, \text{hot})\}, \{(a, \text{reading})\}\}$,
- $\mathbf{L_{wf}}(\mathbf{Q}_3) = \{\{(t, \text{hot})\}\}$.

We define below the preorder corresponding to $\mathbf{L_{wf}}$.

Definition 11. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query and* δ *the weighting on* $\mathbb{E}(\mathbb{T})$ *such that for any* $E \in \mathbb{E}(\mathbb{T})$*:*

$$\delta(E) = \begin{cases} 1 & \text{if } E \cap x = \emptyset \text{ and } \kappa(x_{\downarrow E}) \neq \kappa(x) \\ +\infty & \text{otherwise} \end{cases}$$

$\succeq^f_{\mathbf{Q}}$ *is a preorder on* $\mathbb{E}(\mathbb{T})$ *s.t.* $\forall E, E' \in \mathbb{E}(\mathbb{T})$, $E \succeq^f_{\mathbf{Q}} E'$ *iff* $(\delta(E) = \delta(E') \text{ and } \mathtt{Feat}(E) \subseteq \mathtt{Feat}(E')) \text{ or } \delta(E) < \delta(E')$

We show that $\mathbf{L_{wf}}$ is an instance of the family of CSRs.

Theorem 10. *The following properties hold:*

- *The function* S *mapping every query* $\mathbf{Q}$ *to* $\succeq^f_{\mathbf{Q}}$ *on* $\mathbb{E}(\mathbb{T})$ *is a faithful ranking.*
- *For any query* $\mathbf{Q}$, $\mathbf{L_{wf}}(\mathbf{Q}) = \max(\mathbb{E}(\mathbb{T}), \succeq^f_{\mathbf{Q}})$.

The second explainer, introduced in [15], captures sparsity by selecting explanations of **minimal cardinality**.

Definition 12. *We define* $\mathbf{L}_c$ *as the explainer that, for any query* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *and* $E \in \mathbb{E}(\mathbb{T})$, $E \in \mathbf{L}_c(\mathbf{Q})$ *iff:*

- $\kappa(x_{\downarrow E}) \neq \kappa(x)$,
- $\nexists E' \in \mathbb{E}(\mathbb{T})$ *s.t* $\kappa(x_{\downarrow E'}) \neq \kappa(x)$ *and* $|E'| < |E|$.

Remark: It is easy to show that $\mathbf{L}_c \sqsubseteq \mathbf{L_{wf}}$.

Ex. 1 (Cont) For every $i = 1, 2, 3$, $\mathbf{L}_c(\mathbf{Q}_i) = \mathbf{L_{wf}}(\mathbf{Q}_i)$.

This explainer ranks CSRs according to their cardinality.

Definition 13. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query and* σ *a weighting on* $\mathbb{E}(\mathbb{T})$ *such that for any* $E \in \mathbb{E}(\mathbb{T})$*:*

$$\sigma(E) = \begin{cases} |E| & \text{if } E \cap x = \emptyset \text{ and } \kappa(x_{\downarrow E}) \neq \kappa(x) \\ +\infty & \text{otherwise} \end{cases}$$

We define $\succeq^c_{\mathbf{Q}}$ *as a preorder on* $\mathbb{E}(\mathbb{T})$ *such that:*

$$\forall E, E' \in \mathbb{E}(\mathbb{T}),\ E \succeq^c_{\mathbf{Q}} E' \text{ iff } \sigma(E) \leq \sigma(E').$$

We show that $\mathbf{L}_c$ is an instance of the family of CSRs.

Theorem 11. *The following properties hold:*

- *The function* S *mapping every query* $\mathbf{Q}$ *to* $\succeq^c_{\mathbf{Q}}$ *on* $\mathbb{E}(\mathbb{T})$ *is a faithful ranking.*
- *For any query* $\mathbf{Q}$, $\mathbf{L}_c(\mathbf{Q}) = \max(\mathbb{E}(\mathbb{T}), \succeq^c_{\mathbf{Q}})$.

Existing approaches that compare candidate counterfactuals using a distance measure—including those proposed in [3, 12, 14, 25, 31, 35, 37, 41, 44]—are **all** defined in the following manner.

Definition 14. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query,* $E \in \mathbb{E}(\mathbb{T})$*, and* d *a distance measure on* $\mathbb{F}(\mathbb{T})$*. An explainer based on* d *is a function* $\mathbf{L}_\mathsf{d}$ *such that* $E \in \mathbf{L}_\mathsf{d}(\mathbf{Q})$ *iff:*

- $\kappa(x_{\downarrow E}) \neq \kappa(x)$,
- $\nexists E' \in \mathbb{E}(\mathbb{T}), \kappa(x_{\downarrow E'}) \neq \kappa(x) \wedge \mathsf{d}(x_{\downarrow E'}, x) < \mathsf{d}(x_{\downarrow E}, x)$

Existing distance-based explainers differ in the choice of distance measure d, and may yield explanations that differ from those produced by $\mathbf{L_{wf}}$ and $\mathbf{L}_c$, as illustrated below.

Ex. 1 (Cont) Let $y = ((t, \text{freezing}), (a, \text{climbing}))$. Assume a distance measure d st. $\mathsf{d}(x_2, x_1) < \mathsf{d}(y, x_1)$. Then, $\mathbf{L}_\mathsf{d}(\mathbf{Q}_1) = \{\{(t, \text{mild})\}\}$.

We define the preorders produced by these explainers.

Definition 15. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query,* d *a distance measure on* $\mathbb{F}(\mathbb{T})$*,* γ *a weighting on* $\mathbb{E}(\mathbb{T})$ *s.t.* $\forall E \in \mathbb{E}(\mathbb{T})$*,*

$$\gamma(E) = \begin{cases} \mathsf{d}(x_{\downarrow E}, x) & \text{if } E \cap x = \emptyset \wedge \kappa(x_{\downarrow E}) \neq \kappa(x) \\ +\infty & \text{otherwise} \end{cases}$$

We define $\succeq_{\mathbf{Q},\mathsf{d}}$ *as a preorder on* $\mathbb{E}(\mathbb{T})$ *such that:*

$$\forall E, E' \in \mathbb{E}(\mathbb{T}),\ E \succeq_{\mathbf{Q},\mathsf{d}} E' \text{ iff } \gamma(E) \leq \gamma(E').$$

We show that $\mathbf{L}_\mathsf{d}$ is an instance of the family of CSRs.

Theorem 12. *Let* d *be distance measure.*

- *The function* S *mapping every query* $\mathbf{Q}$ *to* $\succeq_{\mathbf{Q},\mathsf{d}}$ *on* $\mathbb{E}(\mathbb{T})$ *is a faithful ranking.*
- *For any query* $\mathbf{Q}$, $\mathbf{L}_\mathsf{d}(\mathbf{Q}) = \max(\mathbb{E}(\mathbb{T}), \succeq_{\mathbf{Q},\mathsf{d}})$.

In Lewisian counterfactuals (first proposed in the seminal paper [28], then in [1, 48]), the change to the set of variables must be small. This can be modeled by the following function $\mathbf{L}_{\mathsf{d},\tau}$, which is also an instance of the family CSRs (i.e., $\mathbf{L}_{\mathsf{d},\tau} \sqsubseteq \mathtt{cSuf}$).

Definition 16. *Let* $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ *be a query,* $E \in \mathbb{E}(\mathbb{T})$*, and* d *a distance measure on* $\mathbb{F}(\mathbb{T})$*. A threshold explainer based on* (d, τ) *is a function* $\mathbf{L}_{\mathsf{d},\tau}$ *s.t* $E \in \mathbf{L}_{\mathsf{d},\tau}(\mathbf{Q})$ *iff* $\kappa(x_{\downarrow E}) \neq \kappa(x)$ *and* $\mathsf{d}(x_{\downarrow E}, x) < \tau$.

There are three notable works that addressed **necessary reasons**. The two, proposed by [20, 29], introduce functions that generate subset-minimal necessary reasons for instances while violating the axiom of sceptical validity. Their reasons are SNRs when core literals exist or when all features in the theory are binary. In the general case, however, the reasons merely guarantee that some change exists, without specifying which one. In this case, they are of no use to a human user. The third work, by [2], studied global counterfactuals. It proposed some axioms, including Coreness, Success, and Non-Triviality, and defined two explainers, say $\mathbf{L}_1$ and $\mathbf{L}_2$: one that returns a single reason, namely the full set of core literals for the class assigned to the instance, and another that returns non-empty, subset-minimal necessary reasons. Both explainers yield GNRs, i.e., $\mathbf{L}_1 \sqsubseteq \mathtt{gNec}$ and $\mathbf{L}_2 \sqsubseteq \mathtt{gNec}$. Our work is more general. It explores a wider spectrum of counterfactual types—including

GNR, SNR, GSR, SSR, and CSR—and contributes deeper theoretical results, most notably a series of representation theorems grounded in six novel axioms.

A notion of necessary reason was introduced in [13] as the intersection of all partial assignments that guarantee the class of an instance. This reason is unique, may be empty, and its removal does not necessarily change the decision. As such, the corresponding explainer does not provide enough information for the user to understand how to alter the prediction, and thus does not qualify as a counterfactual explanation.

7.2 Properties of Counterfactual (Explainers)

Properties of explanations have been studied in several papers [5, 16, 40, 46, 47] and various properties of couterfactuals have been proposed in [23, 26, 30, 37, 38, 42, 43]. These papers focused on credulous sufficient reasons (CSRs) and discussed various quantitative properties that serve to define functions that select a subset of counterfactuals returned by cSuf. Examples of properties are *Minimality, Similarity, Plausibility, Discriminative power, Actionability, Causality* and *Fairness*. Minimality guarantees a minimal cardinality, similarity a minimal distance from the instance to be explained, whereas plausibility states that the new values in a CSR should be similar to those of instances seen during the training phase [23, 30, 37, 38, 42]. Discriminative power is a subjective property which states that a counterfactual should help in figuring out why a different outcome can be obtained. Actionability guarantees that the features are actionable, i.e., can be mutated. This property is of great importance and may lead to the violation of Success. Causality guarantees that counterfactuals preserve known causal relations between attributes [42]. A counterfactual is fair if it does not refer to protected features [26, 43]. Fairness of explanations is intertwined with fairness of decisions; indeed, the former can be used to evaluate the latter [19].

Properties of sets of counterfactual explanations have also been proposed, namely *Diversity* and *Stability*, called also *Robustness* [22]. Diversity states that chosen counterfactuals should be different, e.g., refer to different attributes [27, 35, 37, 39]. Stability [17] posits that similar instances should receive similar explanations.

Remark: Sections 4 and 5 introduced five broad families of explainers, each producing a specific type of counterfactual (GNR, SNR, GSR, SSR, CSR). The properties presented above **complement** this classification by helping to identify 'reasonable' explainers within each family. For example, one may define an explainer that generates SNRs that are subset-minimal, actionable, diverse, and fair.

Finally, while [4] studied properties of explainers for **'why'** questions and analyzed decision complexity for three **fixed** classifiers, our work focuses on **'why not'** questions. We introduce novel axioms that characterize five families of counterfactual explainers and analyze explanation complexity under **arbitrary** classifiers.

8 COMPUTATIONAL COMPLEXITY

This section analyses two basic computational problems associated with a counterfactual explainer $\mathbf{L}$:

- DecideExp($\mathbf{L}$) is the decision problem whose input is a query $\mathbf{Q}$ and an explanation E, and which determines whether $E \in \mathbf{L}(\mathbf{Q})$.
- FindExp($\mathbf{L}$) takes as input a query $\mathbf{Q}$ and returns an element $E \in \mathbf{L}(\mathbf{Q})$ (or "none" if none exists).

The complexity of these problems depends on the family of possible classifiers κ. To be concrete, we assume that κ is a known boolean formula. If complexity depends on domain size, we also consider larger domains.

Theorem 13. *Consider query $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ where $\mathbb{T}$ is composed of n boolean features and κ is a boolean formula.*

1. *The problem* DecideExp($\mathbf{L}$) *is $O(n)$ when $\mathbf{L}$ is* cSuf, sNec, *or* $\mathbf{L}_{\mathbf{wf}}$.
2. *The problem* DecideExp($\mathbf{L}$) *is co-NP-complete when $\mathbf{L}$ is* gNec, gSuf, sSuf, $\mathbf{L}_{\mathbf{c}}$, $\mathbf{L}_d$.

For domain sizes which are arbitrary, DecideExp($\mathbf{L}$) *is $O(n)$ when $\mathbf{L}$ is* cSuf *or* $\mathbf{L}_{\mathbf{wf}}$. *For queries over non-boolean domains,* DecideExp($\mathbf{L}$) *is co-NP-complete when $\mathbf{L}$ is* sNec.

We study the complexity of generating an explanation.

Theorem 14. *Consider query $\mathbf{Q} = \langle \mathbb{T}, \kappa, x \rangle$ where $\mathbb{T}$ is composed of n boolean features and κ is a boolean formula.*

1. *The problem* FindExp($\mathbf{L}$) *is $O(n)$ when $\mathbf{L}$ is* sSuf.
2. *The problem* FindExp($\mathbf{L}$) *is NP-hard when $\mathbf{L}$ is* gNec, cSuf, sNec, gSuf, $\mathbf{L}_{\mathbf{c}}$, $\mathbf{L}_d$, *or* $\mathbf{L}_{\mathbf{wf}}$. *In each of these cases, excluding $\mathbf{L}_{\mathbf{c}}$ and $\mathbf{L}_d$, it can be solved by at most n calls to a SAT oracle.*

Over non-Boolean domains, FindExp($\mathbf{L}$) *is NP-hard for* sSuf.

Theorems 13 and 14 provide upper bounds on complexity for white-box classifiers expressible as boolean formulas. However, it should be pointed out that for specific types of classifiers, such as decision trees or monotonic functions, for example, certain explainability queries are polynomial-time [4, 11, 18, 21, 32].

9 CONCLUSION

This paper introduces the first axiomatic framework for counterfactual explanations. We define a set of axioms and establish representation theorems that characterize five types of counterfactuals, three of which are novel: SNR, GSR, and SSR. These types fall into two fundamental forms—necessary and sufficient reasons. Necessary reasons are concise and informative but may not always exist, whereas sufficient reasons are guaranteed to exist but can be numerous, especially when feature domains are large. Both forms are generally intractable to compute, except in restricted settings such as Boolean feature spaces.

This work opens several avenues for future research. One direction is to characterize, for each type of counterfactual, the explainers that generate explanations of that type while satisfying established properties such as actionability, minimality, fairness, and diversity. Another direction is to study the robustness of the newly introduced types of counterfactuals.

ACKNOWLEDGMENTS

Our work was supported by the AI Interdisciplinary Institute ANITI, funded by the France 2030 program under the Grant agreement n°ANR-23-IACL-0002, and by the French National Research Agency project ForML ANR-23-CE25-0009.

REFERENCES

[1] Carlos Aguilera-Ventura, Andreas Herzig, Xinghan Liu, and Emiliano Lorini. 2023. Counterfactual Reasoning via Grounded Distance. In *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR*. 2–11.

[2] Leila Amgoud and Jonathan Ben-Naim. 2022. Axiomatic Foundations of Explainability. In *IJCAI*. 636–642.

[3] Gilles Audemard, Jean-Marie Lagniez, and Pierre Marquis. 2024. On the Computation of Contrastive Explanations for Boosted Regression Trees. In *27th European Conference on Artificial Intelligence, ECAI*, Vol. 392. IOS Press, 1083–1091.

[4] Shahaf Bassan, Guy Amir, and Guy Katz. 2024. Local vs. Global Interpretability: A Computational Complexity Perspective. In *ICML (Proceedings of Machine Learning Research, Vol. 235)*. 3133–3167. https://openreview.net/forum?id=veEjiN2w9F

[5] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2023. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery* 37, 5 (2023), 1719–1778.

[6] Ruth Byrne. 2016. Counterfactual Thought. *Annual Review of Psychology* 67 (2016), 135–157.

[7] Ruth Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*. 6276–6282.

[8] Jörg Cassens, Lorenz Habenicht, Julian Blohm, Rebekah Wegener, Joanna Korman, Sangeet Khemlani, Giorgio Gronchi, Ruth M. J. Byrne, Greta Warren, Molly S. Quinn, and Mark T. Keane. 2021. Explanation in Human Thinking. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. cognitivesciencesociety.org.

[9] Jianming Chen, Yawen Wang, Junjie Wang, Xiaofei Xie, Jun Hu, Qing Wang, and Fanjiang Xu. 2025. Understanding Individual Agent Importance in Multi-Agent System via Counterfactual Reasoning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*. AAAI Press, 15785–15794.

[10] Martin C. Cooper and Leila Amgoud. 2023. Abductive Explanations of Classifiers Under Constraints: Complexity and Properties. In *ECAI*, Vol. 372. IOS Press, 469–476.

[11] Martin C. Cooper and João Marques-Silva. 2023. Tractability of explaining classifier decisions. *Artificial Intelligence* 316 (2023), 103841. https://doi.org/10.1016/j.artint.2022.103841

[12] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-Objective Counterfactual Explanations. In *Parallel Problem Solving from Nature – PPSN XVI*. Springer International Publishing, 448–469.

[13] Adnan Darwiche and Auguste Hirth. 2020. On the Reasons Behind Decisions. In *24th European Conference on Artificial Intelligence ECAI*, Vol. 325. IOS Press, 712–720.

[14] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Pai-Shun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Annual Conference on Neural Information Processing Systems, NeurIPS*. 590–601.

[15] Riccardo Guidotti. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining Knowledge Discovery* 38, 5 (2024), 2770–2824. https://doi.org/10.1007/S10618-022-00831-6

[16] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. 2019. Black Box Explanation by Learning Image Exemplars in the Latent Feature Space. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*. 189–205.

[17] Riccardo Guidotti and Salvatore Ruggieri. 2019. On The Stability of Interpretable Models. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8.

[18] Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, Martin C. Cooper, Nicholas Asher, and João Marques-Silva. 2022. Tractable Explanations for d-DNNF Classifiers. In *AAAI*. AAAI Press, 5719–5728. https://doi.org/10.1609/AAAI.V36I5.20514

[19] Alexey Ignatiev, Martin C. Cooper, Mohamed Siala, Emmanuel Hebrard, and João Marques-Silva. 2020. Towards Formal Fairness in Machine Learning. In *CP (Lecture Notes in Computer Science, Vol. 12333)*. 846–867.

[20] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and João Marques-Silva. 2020. From Contrastive to Abductive Explanations and Back Again. In *AIxIA 2020 (LNCS, Vol. 12414)*, Matteo Baldoni and Stefania Bandini (Eds.). Springer, 335–355.

[21] Yacine Izza, Alexey Ignatiev, and João Marques-Silva. 2022. On Tackling Explanation Redundancy in Decision Trees. *J. Artif. Intell. Res.* 75 (2022), 261–321.

[22] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. 2024. Robust Counterfactual Explanations in Machine Learning: A Survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 8086–8094.

[23] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. 2022. Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, Vol. 151. PMLR, 1846–1870.

[24] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *Comput. Surveys* 55, 5 (2022).

[25] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*. 4466–4474.

[26] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *NIPS*. 4066–4076.

[27] Francesco Leofante and Nico Potyka. 204. Promoting Counterfactual Robustness through Diversity. In *AAAI*, Vol. 38(19). 21322–21330.

[28] David Lewis. 1973. *Counterfactuals*. Harvard University Press, Cambridge, MA.

[29] Xinghan Liu and Emiliano Lorini. 2023. A unified logical framework for explanations in classifier systems. *Journal of Logic and Computation* 33, 2 (2023), 485–515.

[30] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable Counterfactual Explanations Guided by Prototypes. In *ECML PKDD*, Vol. 12976. 650–665.

[31] Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Karthikeyan Shanmugam, and Chun-Chen Tu. 2019. Generating Contrastive Explanations with Monotonic Attribute Functions. *CoRR* (2019). http://arxiv.org/abs/1905.12698

[32] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. 2021. Explanations for Monotonic Classifiers. In *ICML 2021 (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 7469–7479.

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

[34] C. Molnar. 2020. *Interpretable Machine Learning*. Lulu.com. https://books.google.fr/books?id=RHjTxgEACAAJ

[35] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, 607–617.

[36] Erfaun Noorani, Pasan Dissanayake, Faisal Hamman, and Sanghamitra Dutta. 2025. Counterfactual Explanations for Model Ensembles Using Entropic Risk Measures. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS*, Sanmay Das, Ann Nowé, and Yevgeniy Vorobeychik (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 1566–1575.

[37] Barry Smyth and Mark T. Keane. 2022. A Few Good Counterfactuals: Generating Interpretable, Plausible and Diverse Counterfactual Explanations. In *ICCBR (LNCS, Vol. 13405)*, Mark T. Keane and Nirmalie Wiratunga (Eds.). Springer, 18–32.

[38] Asterios Tsiourvas, Wei Sun, and Georgia Perakis. 2024. Manifold-Aligned Counterfactual Explanations for Neural Networks. In *International Conference on Artificial Intelligence and Statistics (PMLR, Vol. 238)*. 3763–3771.

[39] Stratis Tsirtsis and Manuel Gomez Rodriguez. 2020. Decisions, Counterfactual Explanations and Strategic Behavior. In *NeurIPS*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[40] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *Comput. Surveys* 56, 12 (2024), 312:1–312:42.

[41] Sahil Verma, John P. Dickerson, and Keegan Hines. 2021. Counterfactual Explanations for Machine Learning: Challenges Revisited. *CoRR* abs/2106.07756 (2021). https://arxiv.org/abs/2106.07756

[42] Sahil Verma, Keegan Hines, and John P. Dickerson. 2021. Amortized Generation of Sequential Counterfactual Explanations for Black-box Models. *CoRR* abs/2106.03962 (2021). https://arxiv.org/abs/2106.03962

[43] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. 2022. On the Fairness of Causal Algorithmic Recourse. arXiv:2010.06529

[44] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* abs/1711.00399 (2017).

[45] Greta Warren, Eoin Delaney, Christophe Guéret, and Mark T. Keane. 2024. Explaining Multiple Instances Counterfactually:User Tests of Group-Counterfactuals for XAI. In *Proceedings of the 32nd International Conference on Case-Based Reasoning Research and Development, ICCBR*, Vol. 14775. 206–222.

[46] Greta Warren, Barry Smyth, and Mark T. Keane. 2022. Better Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI). In *Case-Based Reasoning Research and Development - 30th International Conference, ICCBR (Lecture Notes in Computer Science, Vol. 13405)*. Springer, 63–78.

[47] Michael Winikoff, John Thangarajah, and Sebastian Rodriguez. 2025. A Scoresheet for Explainable AI. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS*. 2171–2180.

[48] James Woodward and Christopher Hitchcock. 2003. Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs* 37, 1 (2003), 1–24.