Explaining with reasons: from Aristotle to machine learning classifiers

Brian Hill, Francesca Poggiolesi

Explaining with reasons: from Aristotle to machine learning classifiers

Brian Hill and Francesca Poggiolesi

Abstract

Explanations, and in particular explanations which provide the reasons why their conclusion is true, are a central object in a range of fields. On the one hand, there is a long and illustrious philosophical tradition, which starts from Aristotle, and passes through scholars such as Leibniz, Bolzano and Frege, that give pride of place to this type of explanations, and is rich with brilliant and profound intuitions. Recently, Poggiolesi (2024a) has formalized ideas coming from this tradition using logical tools of proof theory. On the other hand, recent work has focused on Boolean circuits that compile some common machine learning classifiers and have the same input-output behavior. In this framework, Darwiche and Hirth (2023) have proposed a theory for unveiling the reasons behind the decisions made by Boolean classifiers, and they have studied their theoretical implications. In this paper, we uncover the deep links behind these two trends, demonstrating that the proof-theoretic tools introduced by Poggiolesi provide reasons for decisions, in the sense of Darwiche and Hirth (2023). We discuss the conceptual as well as the technical significance of this result.

1 Introduction

Explanations are a central object of study in a range of fields. In particular, in the last ten years or so, explanations that display the reasons why a certain conclusion is true, or why a certain phenomenon occurs, have become a flourishing and thriving object of research in several disciplines. In philosophy, for example, there is a long and illustrious tradition - starting with Aristotle and passing by scholars such as Leibniz, Bolzano or Frege[1] - that, although often neglected for several decades, is now witnessing a renewed interest. This tradition gives pride of place to explanations that play the same role in conceptual sciences, like mathematics, as causal explanations do in the empirical sciences: as the latter explain by displaying the causes, the former explain via the reasons.[2] Recently, Poggiolesi (2024a) has formalized ideas coming from this tradition using logical tools from proof theory. She has thus introduced a sequent calculus with explanatory rules, namely rules where not only the conclusion follows from the premise(s), but where the premise(s) represent the reasons why the conclusion is true.

Another notable field where explanations with reasons have been receiving an increasing amount of attention and interest is computer science, and in particular machine learning classifiers. A recent research direction has aimed to show that some common machine learning classifiers can be represented by propositional formulas and classifications of instances can be analysed in terms of classical logical consequence. In this context, Darwiche and Hirth

[1] E.g., see Betti (2010) or Detlefsen (1988).

[2] E.g. see Mancosu et al. (2023); Poggiolesi and Genco (2023).

(2023) define the idea of sufficient reasons for a decision, which correspond to its potential *explicanda*,[3] and relate it to prime implicants, as well as a range of related concepts.

The notion of explanation with reasons thus emerges as a central topic in two distinct fields, linked to different literatures, and using different techniques to develop different approaches. Is there a connection, beyond the apparent similarity in object? The current paper provides an answer to this question, by mapping notions used in machine learning into the framework for understanding explanations developed in philosophy. More precisely, we show that the explanatory tools introduced by Poggiolesi (2024a) naturally provide all and only the sufficient reasons for a decision, in the sense of Darwiche and Hirth (2023). This result is of double conceptual significance. On the one hand, it shows that the philosophical approach to explanation via explanatory rules has applications in practical and significant cases, where it naturally captures pre-philosophical intuitions. On the other hand, it connects the notion of sufficient reason introduced in AI to a long philosophical tradition, suggesting that there exists a sort of *fil rouge* going from Aristotle to the present day, confirming the deepness and interest of the concepts at issue. Our result is also of formal import, insofar as the computation of the set of sufficient reasons for a decision has long been an open problem (Coudert and Madre, 1993; Shih et al., 2018). The result shows that the explanatory framework developed in philosophy naturally provides a novel and simple solution to this problem, which may be of use in the study of as-yet open generalisations.

The paper is structured as follows. In *Section 2* we introduce the concepts of sufficient and complete reason as defined by Shih et al. (2020), whilst in *Section 4* we clarify the main ideas behind the explanatory sequent calculus of Poggiolesi (2024a) and a generalized version of it. *Section 4* will be used to illustrate how we can use the explanatory calculus to compute sufficient reasons behind decisions made by boolean classifiers, whilst in *Section 5* we will prove that explanatory calculi can compute the complete reasons behind decisions made by classifiers. We finally close with some concluding remarks in *Section 6*.

2 Sufficient and complete reasons

We use this section to introduce and clarify the key notions proper of the account developed by Darwiche and Hirth (2023), Shih et al. (2020). To do so, we proceed by means of the following example. Consider a classifier $\mathcal{C}$ which admits an applicant just in case she has passed the entrance exam and either she is not a first time applicant, or she has work experience, or she has a high GPA. Consider the language of classical propositional language defined as follows.

Definition 2.1. The language of classical propositional logic, $\mathcal{L}$, is composed by: atomic formulas (p_0, p_1, p_2, ...), logical connectives ($\neg$, $\wedge$, $\vee$), and parentheses: (,). *Literals* are either atomic formulas or their negations, whilst *extended literals* are formulas of the form: $\overbrace{\neg \dots \neg}^{n} p$ for $n \geqslant 0$ and atomic formula p. Extended literals will be denoted by $l_1, l_2, \dots$. We take the symbols $\top, \bot$ and $\rightarrow$ to be defined as usual. The set of well-defined formulas, $\mathcal{WF}$, is constructed in the standard way. Finally, for the sake of brevity, the symbol $\circ$ will be used to denote either a conjunction or a disjunction.

As it will become clear later on, it is convenient to use the metalinguistic symbol of *converse of a formula*, which is defined as follows.

[3] Also known as PI-explanations Shih et al. (2018), or abductive explanations Ignatiev et al. (2020). Further extensions of this work can be found in e.g. Liu and Lorini (2023, 2022).

Definition 2.2. The converse of a formula A, written $A^{\perp}$, is defined as follows:

$$A^{\perp} = \begin{cases} \neg^{n-1}E, & if\ A = \neg^{n}B\ and\ \text{n}\ is\ odd \\ \neg^{n+1}E, & if\ A = \neg^{n}B\ and\ \text{n}\ is\ even \end{cases}$$

where the main connective in B is not a negation, $n \geqslant 0$ and 0 is taken to be an even number.

Suppose to use the following denotations:

- e denotes "applicant has passed the entrance exam,"
- f denotes "applicant is at her the first attempt,"
- w denotes "applicant has work experience,"
- g denotes "applicant has a high GPA."

Following Darwiche and Hirth (2023), we can represent classifier $\mathcal{C}$ by the following formula A of the classical propositional language, $A = e \wedge (\neg f \vee w \vee g)$, where e, f, w, g are literals.

Consider now the two applicants Greg and Susan. Greg and Susan have both passed the entrance exam, but whilst Greg has already applied and has work experience, Susan has work experience and a high GPA. Their features can be gathered in the following two sets (of literals) α_g and α_s, respectively, and can be formalized as follows $\alpha_g = \{e, \neg f, w, \neg g\}$ and $\alpha_s = \{e, f, w, g\}$. We call both α_g and α_s instances, where instances are defined as follows.

Definition 2.3. An *instance* α of a formula A is a consistent set of literals such that, for every literal l appearing in A, either l or $l^{\perp}$ is in α. We sometimes refer to the literals in α as the *characteristics* of the instance.

Grege and Susan are both admitted by classifier $\mathcal{C}$, as the admission is granted in case the formula which represents the classifier logically follows from the instance that corresponds to the applicant. It is easy to verify that we have both $\alpha_g \models A$, as well as $\alpha_s \models A$.

In more general terms, we use $A(\alpha)$ to denote the decision (0 or 1) of classifier A on instance α: this corresponds to $\alpha \models A$ if, and only if, $A(\alpha) = 1$, and $\alpha \models \neg A$ if, and only if, $A(\alpha) = 0$. We also define A_{α} in the following way:

$$A_{\alpha} = \begin{cases} A, & \text{if } A(\alpha) = 1 \\ \neg A, & \text{if } A(\alpha) = 0 \end{cases}$$

Note that $\alpha \models A_{\alpha}$ and $A(\alpha) = A(\beta)$ if, and only if, $A_{\alpha} = A_{\beta}$.

Hence, both Greg and Susan have been admitted by classifier $\mathcal{C}$; however, it seems that it is so for different reasons. The main goal of Darwiche and Hirth (2023), Shih et al. (2020) is to explain the decisions made by a classifier on specific instances by way of providing various insights into what motivated these decisions. In particular, they propose the captivating notion of *sufficient reason* which is properly defined as follows via the notion of *implicant*.

Definition 2.4. A *term* is a consistent set of literals. Terms will be denoted by $\tau_1, \tau_2, \tau_3, ...$ Terms may also be called *properties* (of instances). A term τ_i subsumes a term τ_j if, and only if, $\tau_i \subseteq \tau_j$.

Definition 2.5. An *implicant* τ of a propositional formula A is a term that satisfies A, namely $\tau \models A$. A *prime implicant* is an implicant that is not subsumed by any other implicant.

Definition 2.6. A *sufficient reason* for decision $A(\alpha)$ is a property of instance α that is also a prime implicant of A_α.

A sufficient reason identifies characteristics of an instance that justify the decision: the decision will remain the same even if other characteristics of the instance were different. Moreover, a sufficient reason is minimal as none of its strict subsets can justify the decision.

Let us go back to the classifier $\mathcal{C}$ which admits an applicant just in case she has passed the entrance exam and either she has already applied or she has work experience or she has a high GPA. According to what has just been said, classifier $\mathcal{C}$ admits Greg for the following sufficient reasons:

- Passed the entrance exam and is not a first time applicant, $\tau_1 = \{e, \neg f\}$.
- Passed the entrance exam and has work experience, $\tau_2 = \{e, w\}$.

Since Greg passed the entrance exam and has applied before, he will be admitted even if his other characteristics were different. Similarly, since Greg passed the entrance exam and has work experience, he will be admitted even if his other characteristics were different. Moreover, in both cases reasons are minimal: if any of their features is omitted, the decision is not longer the same.

Note that classifier $\mathcal{C}$ also admits Susan for the following sufficient reasons:

- Passed the entrance exam and has work experience, $\tau_1 = \{e, w\}$.
- Passed the entrance exam and has a high GPA, $\tau_2 = \{e, g\}$.

Since Susan passed the entrance exam and has work experience, or since Susan passed the entrance exam and she has a high GPA, she will be admitted. This is so even if her other characteristics were different and again these sets are minimal.

Definition 2.7. The *complete reason* for a decision is the set of all sufficient reasons.

A decision may have multiple sufficient reasons, sometimes an exponential number of them, e.g. see Darwiche (2018). As a result the enumeration of all sufficient reasons, i.e. the enumeration of the complete reasons, is a problem that has been treated in several papers, e.g. see Coudert and Madre (1993) and Shih et al. (2018). The algorithms proposed over the time also had to face another difficulty which can be easily explained by means of the following example. Consider the classifier $\mathcal{C}'$ represented by the formula $A' = (p \wedge q) \vee (r \wedge \neg q)$ and the instance $\alpha' = \{p, q, r\}$. It can be straightforwardly checked that the decision $A'(\alpha')$ has two sufficient reasons, namely $\tau = \{p, q\}$, but also $\tau' = \{p, r\}$ (because of the law of excluded middle). Algorithms proposed in the above mentioned literature may miss the sufficient reason $\tau' = \{p, r\}$ and therefore be incomplete. Darwiche and Hirth (2023) and Shih et al. (2020) have recently fixed this problem by constructing algorithms that provide in linear time the set of complete reasons; however, as the discussion shows, this is no trivial solution.

3 Explanatory sequent calculus

Extending ideas introduced in Genco (2021) and Poggiolesi (2018), Poggiolesi (2024a) mobilizes the resources of proof theory to formalize the notion of explanation from reasons. Two main ideas motivate this formalization. The first relates to a remark that is both ancient and central in the philosophical literature on explanations, and which consists at looking to explanations as deductive *arguments* that, starting from true premisses - be they the causes or the reasons - explain a certain conclusion.[4] Of course not any deductive argument constitutes an explanation, but some of them do, namely those which have an explanatory power. The perspective that is adopted in Poggiolesi (2024a) consists in a formalization of this central idea along the following lines: explanations can be seen as *proofs* which, starting from true premisses, the reasons, not only prove that a certain conclusion is true, but also explain why it is such. This perspective naturally arises from the observation that proofs are deductive arguments; moreover, it is supported by the fact that notable explanations from reasons that can be found in mathematics actually are proofs of mathematical theorems, which show why those theorems are true.[5] Since proofs are standardly formalized in logic by means of *derivations*, explanations from reasons will be formalized as a special type of derivations, namely a metalinguistic relation called *formal explanation*. As derivations are introduced via inferential rules, formal explanations will be introduced via *explanatory rules*, namely rules where not only is the conclusion inferable from their premise(s), but also such that the premisses are the reasons why the conclusion is true.

Let us now move to the second main idea linked to the formalization of explanations proposed by Poggiolesi (2024a). The second idea relates to the observation that in many examples of explanation from reasons, reasons and conclusions, once formalized in a formal language, correspond to formulas related to each other by elements that occur inside the formulas themselves.[6] Consider the sentence "any bachelor is an unmarried man;" the intuitive reasons which explain this truth are "any bachelor is unmarried" and "any bachelor is a man." Formalizing them we obtain $\forall x(Bx \rightarrow Ux \wedge Mx)$, and $\forall x(Bx \rightarrow Ux)$, $\forall x(Bx \rightarrow Mx)$, respectively. The link between these formulas occurs deep inside the formulas themselves: in particular, the connective $\wedge$ inside $\forall x(Bx \rightarrow Ux \wedge Mx)$ is broken into two and thus give rise to both $\forall x(Bx \rightarrow Ux)$ and $\forall x(Bx \rightarrow Mx)$. However, as already said, this is no isolate case. Another example is obtained by considering the sentence"zero or the successor of any natural number it is itself a natural number." The reasons which explain this truth are "zero is a natural number" and "the successor of any natural number is a natural number." Formalizing them we obtain $\forall x((Zx \vee SNx) \rightarrow Nx)$, and $\forall x(Zx \rightarrow Nx)$, $\forall x(SNx \rightarrow Nx)$. Once again, the link between these formulas occurs deep inside the formulas themselves: in particular, this time it is the connective $\vee$ which is broken and gives rise to the reasons of the conclusion.[7] Similar examples also arise at the propositional level. Consider the sentence "if it is a triangle, then it has three sides, and the sum of its angles is 180° ;" the reasons which explain its truth are "if it is a triangle, then it has three sides" and "if it is a triangle, then the sum of its angles is 180°." Formalizing them we obtain $t \rightarrow s \wedge a$, and $t \rightarrow s$, $t \rightarrow a$, respectively. Once more the link between these formulas occurs deep inside the formulas

[4] E.g. see Aristotle (1993); Hempel (1965, 1942).

[5] E.g. see Betti (2010); Mancosu et al. (2023); Poggiolesi and Genco (2023).

[6] This becomes especially evident when considering formulas from first-order logic, which is indeed the framework where Poggiolesi (2024a) works. Although for the rest of the paper we will remain at the propositional level, as this is the level required to establish a connection with the work of Darwiche and Hirth (2023), we will now pick example involving quantifiers as they are illustrative.

[7] Many other example involving the quantifiers can be found in Poggiolesi (2024b); Poggiolesi and Genco (2023).

themselves.

To deal with this kind of cases, Poggiolesi introduces the notions of *context* and *formula in a context* that allow to focus on a particular part of the formula at issue.[8] For example consider again the formula $\forall x((Zx \vee SNx) \rightarrow Nx)$ and suppose we want to focus on a particular part of it, say $Zx \wedge SNx$. We denote this fact by rewriting $\forall x((Zx \vee SNx) \rightarrow Nx)$ as $C[Zx \wedge SNx]$, where $C[.]$ is the context and $Zx \wedge SNx$ is the formula in the context $C[.]$.

Note that when working in an explanatory framework, negation needs to be taken into account with special attention.[9] This is also true for the notion of context, as can be clearly seen in the following example, concerning the formulas $\neg(p \vee q)$ and $\neg(\neg p \vee \neg q)$. Whilst the reasons of $\neg(p \vee q)$ amount to the formulas $\neg p, \neg q$, the reasons of $\neg(\neg p \vee \neg q)$ are p, q. However, if we take a negation (or any odd number of consecutive negations) in front of a disjunction to be a context, and the reasons of a disjunction to be its disjuncts, we would get that the reasons for $\neg(p \vee q)$ are indeed $\neg p, \neg q$, whilst the reasons for $\neg(\neg p \vee \neg q)$ are $\neg\neg p, \neg\neg q$, contrary to what has just been noted. To avoid such undesirable cases, we define contexts only on an even consecutive number of negations, and we will treat the negation of a disjunction with special rules that involve the notion of *converse of a formula* that we have introduced in the previous section.

Definition 3.1. The set Co of contexts is inductively defined in the following way:

- $[.] \in Co$,
- if $C[.] \in Co$, then $\neg\neg C[.]$, $D \circ C[.]$, $C[.] \circ D \in Co$,
- if $C[.] \in Co$ and $C[.] \neq \overbrace{\neg...\neg}^{2n}[.]$, where $n \geqslant 0$, then $\neg C[.] \in Co$.

Definition 3.2. For all contexts $C[.]$, and formulas A, we define $C[A]$, a formula in a context, as follows:

- if $C[.] = [.]$, then $C[A] = A$,
- if $C[.] = \neg\neg D[.]$, then $C[A] = \neg\neg D[A]$,
- if $C[.] = D' \circ D[.]$, $D[.] \circ D'$, $\neg D[.]$, then $C[A] = D' \circ D[A]$, $D[A] \circ D'$, $\neg D[A]$, respectively.

Once formulas are considered in contexts, they will naturally have a polarity which is either positive or negative and that is defined as standard, e.g. see Troelstra and Schwichtenberg (1996).

Definition 3.3. We define the set of contexts with positive $\mathcal{P}$ and negative polarities $\mathcal{N}$ simultaneously by an inductive definition given by the three clauses (i)-(iii) below.

- $[.] \in \mathcal{P}$;

if $B^+ \in \mathcal{P}$, $B^- \in \mathcal{N}$, and A is any formula, then:

[8] Most likely the lack of any reference to the notion of context in the account developed by Darwiche and Hirth (2023) is motivated by the fact that this account focuses on propositional formula, whilst it is when quantifiers are introduced that the need of context becomes particularly evident. However, this difference between Darwiche and Hirth (2023) and Poggiolesi (2024a) is worth further investigation.

[9] E.g., see Poggiolesi (2016b, 2022).

(ii) $\neg B^-$, $A \wedge B^+$, $B^+ \wedge A$, $A \vee B^+$, $B^+ \vee A \in \mathcal{P}$.

(iii) $\neg B^+$, $A \wedge B^-$, $B^- \wedge A$, $A \vee B^-$, $B^- \vee A \in \mathcal{N}$

whenever these objects are in Co. We say that a formula A is positive (resp. negative) in a context $C[A]$ if $C[.] \in \mathcal{P}$ (resp. $C[.] \in \mathcal{N}$).

Summing up, we have introduced the two main ideas that characterize Poggiolesi's formalism. The first idea amounts at looking at explanations as special types of proofs, whilst the second amounts to the exigence of working deep inside formula. Putting them together, we have that Poggiolesi (2024a) introduces explanatory rules that deal with formulas in contexts. In other words, explanatory rules will have the form of *deep inferences*, namely a recently introduced variation of the sequents calculus (e.g. see Brünnler (2004); Guglielmi and Bruscoli (2009); Pimentel et al. (2019)) where rules operate deep inside formulas. Although the literature on deep inferences has been motivated by cornerstone results of structural proof theory, in this context they reveal a profound philosophical significance.

As already stated, explanatory rules are such that their premises are the reasons why their conclusion is true. More specifically, explanatory rules provide the *total* reasons of why the conclusion is true, where total reasons typically amount to the multiset of all, and only, those formulas, each of which contributes to explain another (see also Schaffer (2016)). Note that the notion of *total reasons* motivates a distinction that we illustrate with the following example. Jane has a brother, Billy, and a sister, Suzy. Jane also has a niece. Thus the reason why Jane has a niece is that her sister has a girl. Indeed a niece is the girl of someone's brother or someone's sister and Suzy, Jane's sister, has a girl. Jane's brother could have had a girl, but he does not. Hence Jane's brother having a girl is merely a potential reason of why Jane has a niece. Potential reasons are also central for total explanations: if Jane's brother had had a girl, this would have been part of the total explanation of why Jane has a niece. We rephrase this distinction between reasons and potential reasons as the one between reasons and *conditions.* So, for example, we will say that under the condition that Jane's brother does not have a girl, the total reason why Jane has a niece is that her sister has a child. We will indicate the distinction between conditions and reasons with a vertical bar: formulas lying at the left of the vertical bar are conditions, whilst formulas at the right of the vertical bar are reasons.

We now have all the elements to introduce sequents for the explanatory calculus **EC**.

Definition 3.4. For any α, A, and multiset $X = \{B_1, ..., B_n\}$, such that α is an instance of A, but also an instance of each B_i, $1 \leqslant i \leqslant n$, we have that:

- $\alpha \Rightarrow A$ is a sequent ;
- $X \mid \alpha \Rightarrow A$ is a sequent,

Definition 3.5. The interpretation τ of a sequent is the following:

- $(\alpha \Rightarrow A)^\tau := \bigwedge \alpha \rightarrow A$;
- $(X \mid \alpha \Rightarrow A)^\tau := (\bigwedge X^\perp \wedge \bigwedge \alpha) \rightarrow A$,

where if $X = \{B_1, ..., B_n\}$, then $\bigwedge X^\perp = B_1^\perp \wedge ... \wedge B_n^\perp$.

By using sequents as defined above, we introduce the explanatory calculus **EC** in Figure1. Note that **EC** results from a modification, but also a generalization, of the ideas conveyed

Figure 1: Explanatory propositional calculus (EC).

Axioms

$$X \mid \alpha, p \Rightarrow p \qquad\qquad X \mid \alpha, \neg p \Rightarrow \neg p$$

Explanatory Rules

$$\frac{X \mid \alpha \Rightarrow C[A]}{X \mid \alpha \Rightarrow C[\neg\neg A]}\ {}^{\neg\neg}$$

$$\frac{X \mid \alpha \Rightarrow C[A] \quad X \mid \alpha \Rightarrow C[B]}{X \mid \alpha \Rightarrow C[A \circ B]}\ {}^{\circ_1} \qquad \frac{X, C[A_j] \mid \alpha \Rightarrow C[A_i]}{X \mid \alpha \Rightarrow C[A_1 \circ A_2]}\ {}^{\circ_2}$$

$$\frac{X \mid \alpha \Rightarrow C[A^{\perp}] \quad X \mid \alpha \Rightarrow C[B^{\perp}]}{X \mid \alpha \Rightarrow C[\neg(A \circ B)]}\ {}^{\neg\circ_1} \qquad \frac{X, C[A_j^{\perp}] \mid \alpha \Rightarrow C[A_i^{\perp}]}{X \mid \alpha \Rightarrow C[\neg(A_1 \circ A_2)]}\ {}^{\neg\circ_2}$$

where both $i, j = \{1, 2\}$ and $j \neq i$.
We assume explanatory rules not to distinguish between formulas which are equivalent. by associativity and commutativity of conjunction and disjunction. Moreover, the application of the explanatory rules is conditioned by Definition 3.6.

in Poggiolesi (2024a). In particular, (i) whilst Poggiolesi introduces explanatory rules for first-order formulas, here we limit ourselves to propositional formulas; (ii) secondly, whilst in the sequents used by Poggiolesi, antecedents and consequents of sequents are multiset of formulas, in the present case, the antecedent is an instance, and the consequent is a single formula; (iii) thirdly, whilst Poggiolesi added explanatory rules to the standard sequent calculus for first-order classical logic, here, in a way similar to that proposed by Genco (2021), explanatory rules plus axioms constitute a calculus on its own. (iv) Finally, in the present work conditions are generalized to multisets of formulas.

Definition 3.6. We assume the application of explanatory propositional rules[10] to obey the following restrictions:

- rule $\circ_2$ can be applied on a formula of the form $C[A \circ B]$ if: $\begin{cases} C \in \mathcal{P} \text{ and } \circ = \vee, \text{or} \\ C \in \mathcal{N} \text{ and } \circ = \wedge. \end{cases}$
- rule $\neg\circ_2$ can be applied on a formula of the form $C[\neg(A \circ B)]$ if: $\begin{cases} C \in \mathcal{P} \text{ and } \circ = \wedge, \text{or} \\ C \in \mathcal{N} \text{ and } \circ = \vee. \end{cases}$

We now comment on the rules of the explanatory sequent calculus. Each of these rules is supposed to capture cases where the premisses are the total reasons for the conclusions[11] Some examples are clear: for instance p and q are clearly the reasons for $p \wedge q$; and rule $\circ_1$ reflects this. Let us then dwell on the less obvious and more novel cases. First of all, note that there is no rule for single negation. This is because explanations notoriously go from (potentially) true formulas to (potentially) true formulas; there can thus be no rule which acts, as in the case of the rule for negation in the standard sequent calculus, by shifting

[10] Reading the rules bottom-up.
[11] Note that in Poggiolesi (2024a), after conditions for certain formulas to count as the total reasons of another are given, this is proved to be the case.

formulas from one side of the sequent to another. In other words, one cannot explain the truth of $\neg A$, from the falsity of A. Instead negation is spread over the other connectives: either it is analyzed when it is double, or when it is in front of conjunction and disjunction. Note that, for reasons mentioned above (when introducing contexts) and which are discussed in Poggiolesi (2016a), the connective of negation must be carefully treated in an explanatory context; this is why the converse of a formula (see Definition 2.2) is used in the rules $\neg\circ_1$ and $\neg\circ_2$.[12]

Let us now turn to those rules that do not involve conditions: i.e. $\neg\neg$, $\circ_1$ and $\neg\circ_1$. Each of them stands as a straightforward generalization of standard rules concerning classical connectives, allowing them to apply deep inside formulas. This is so because these rules are not merely intended to be simple inferential rules but explanatory rules, i.e. rules that provide the (total) reason(s) why their conclusion is true. The relation between reason(s) and conclusion might hold in virtue of elements that lie inside formulas, so the rules need to reflect this possibility.

Let us now move to the rules which involves conditions, namely the rules $\circ_2$, $\neg\circ_2$. These rules naturally emerge for total explanations, i.e. explanations where all the reasons why a conclusion is true need to be evoked. In this setting, conditions need to be mentioned to prevent equivocation between total and partial explanations (e.g. see Poggiolesi (2016a)). Consider the example: John got into the University, and he is rich or he passed the entrance exam. Suppose that in fact John got into the University, he is rich, but he did not pass the entrance exam. In this example, the explanation why it is true that John got into the University, and he is rich or he passed the entrance exam is that John got into University and he is rich. However, if nothing is said about the passing exam, the explanation remains ambiguous: it is indeed unclear whether the explanandum is true also because John got into University and passed the entrance exam. Conditions allow disambiguation of the explanation. Thus we say that, under the condition that it is not the case that John got into University and passed the entrance exam, it is true that John got into the University, and he is rich or he passed the entrance exam, because John got into University and he is rich. On formal terms, let us denote the sentence "John gets into the University, and he is either rich or it has passed the entrance exam," with the formula $p \wedge (q \vee r)$. Let us apply on this formula, focussing on the disjunction, the following instance of the rule $\circ_2$, we get:

$$\frac{\Rightarrow p \wedge q \mid \Rightarrow p \wedge r}{\Rightarrow p \wedge (q \vee r)}$$

The rule can be applied since $\circ$ is a disjunction with a positive polarity (see Definition 3.6). Thanks to the rule $\circ_2$, we can explain the formula $p \wedge (q \vee r)$ by the formula $p \wedge r$, which represents the total reason why it is true under the condition that the formula $p \wedge q$ does not hold. The rule matches what we have just been discussing and thus stands as a an adequate instance of the rule.

Applications of rules deep inside formulas with conditions involve some limitations: these limitations serve to preserve an adequate notion of explanation. For example, consider the formula $p \rightarrow (q \wedge r)$. One cannot apply the rule $\circ_2$ on the conjunctive of this formula, since the conjunction occurs with positive polarity. On the other hand, this limitation is recommended. Indeed, the rule might have the following instance:

[12]Indeed thanks to the notion of converse and following the rule $\neg\circ_1$, we have that the reasons why the formula $\neg(p \vee q)$ is true are $\neg p$ and $\neg q$, whilst the reasons why the formula $\neg(\neg p \vee \neg q)$ is true are p and q. This result fits with the example discussed above

$$\frac{p \to q \mid\Rightarrow p \to r}{\Rightarrow p \to (q \wedge r)}$$

and it is easy to check that the formula $(\Rightarrow p \to (q \wedge r))^{\tau}$ is not even derivable from the formula $(p \to q \mid\Rightarrow p \to r)^{\tau}$ (see Definition 3.5 for the interpretation of sequent).

Finally note that explanatory rules do not distinguish between formulas that are equivalent by associativity and commutativity of conjunction and disjunction. This feature contributes for explanatory rules to have a hyperintensional flavor, which is in line with the fact that explanation is an hyperintensional notion, e.g. see Berto and Nolan (2023); Leitgeb (2019).

Definition 3.7. Let $\vdash_{EC} X \mid \alpha \Rightarrow A$ denote that there exists a derivation of the sequent $X \mid \alpha \Rightarrow A$ in the explanatory calculus EC.

Theorem 3.8. *For any X, α and A, if $\vdash_{EC} X \mid \alpha \Rightarrow A$, then $\models (X \mid \alpha \Rightarrow A)^{\tau}$.*

Proof. Straightforward, given Theorem 5.7 in Poggiolesi (2024a).

□

Theorem 3.9. *For any α and A, if $\alpha \models A$, then $\vdash_{EC} \alpha \Rightarrow A$.*

Proof. Suppose $\alpha \Rightarrow A$ is not a derivable sequent. This means that whatever sequence of the explanatory rules $\neg\neg, \circ_1, \neg\circ_1$ we apply on A, we will end up with leafs of the form $\alpha, p \Rightarrow \neg p$, or $\alpha, \neg p \Rightarrow p$. This means that if we assign evaluation 1 to the literals on the left side of the sequent, the literals on the right side of the sequents will have evaluation 0. Since, as it has been shown in Theorem 5.7 in Poggiolesi (2024a), explanatory rules preserve the evaluation 0, then A will have evaluation 0 as well. Hence we have found an evaluation witnessing that A is not a logical consequence of α.

□

4 From the sequent calculus to sufficient reasons

We now describe how to use the explanatory calculus to extract sufficient (but also, as we will show, complete) reasons out of a decision. To do so, we first of all give instructions on how to define a comprehensive derivation of a sequent $\alpha \Rightarrow A$.

Definition 4.1. The depth of a subformula in a formula is the number of connectives it is nested into. So, for the example, the depth of $B \vee C$ in the formula $E \wedge (B \vee C)$ is 1; whilst the depth of $B \vee C$ in the formula $(E \vee ((F \wedge (B \vee C)) \wedge (G \wedge H))) \wedge (G \vee H \vee R)$ is 4.

Definition 4.2. Consider the sequent $\alpha \Rightarrow A$. A derivation d of $\alpha \Rightarrow A$ is *comprehensive* if (reading the derivation top-down):

1. For every sequent of the form $X \mid \alpha \Rightarrow A'$ appearing in d and every $B_i \in X$, $\alpha \Rightarrow B_i$ is neither an axiom nor derivable via explanatory rules.

2. If $R \circ' \overbrace{\neg \ldots \neg}^{m}(B_1 \circ'' B_2)$ is a subformula of A' with $\alpha \Rightarrow A'$ appearing in d, for $\circ', \circ'' \in \{\vee, \wedge\}$, with $\circ' \neq \circ''$ if $m = 2n$, and $\circ' = \circ''$ if $m = 2n + 1$, then:

   (a) if $(B_1 \circ'' B_2) \neq (l \vee \neg l), (l \wedge \neg l)$ for any extended literal l, then any rule applied on $\circ'$ in this subformula is higher in d than every rule applied on $\circ''$;

(b) $(B_1 \circ'' B_2) = (l \vee \neg l), (l \wedge \neg l)$, for any extended literal l, then any rule applied on $\circ''$ in this subformula is higher in d than every rule applied on $\circ'$.

3. If $B_1 \circ' ... \circ' B_r$ is a subformula of A' with $\alpha \Rightarrow A'$ appearing in d, where each $B_i = \overbrace{\neg...\neg}^{2n}(l_1 \circ'' ... \circ'' l_{r_i})$ or $\overbrace{\neg...\neg}^{2n+1}(l'_1 \circ'' ... \circ'' l'_{r'_i})$ for $\circ', \circ'' \in \{\vee, \wedge\}$, $\circ' \neq \circ''$, then any application of rules $\circ_2$, $\neg\circ_2$ in d with main connective in this subformula occurs below applications of other rules with main connective in this subformula.

Since comprehensive derivations are just derivations with specifications on the order of application of rules, by Theorem 3.9, whenever $\alpha \models A$, there exists a comprehensive derivation of $\alpha \Rightarrow A$. Moreover, comprehensive derivations can easily be constructed using a version of a standard proof search algorithm: details are provided in Appendix A. The constructed derivation is guaranteed to have proof height (maximal length of a branch from root to leaf) at most the complexity of the classifier A. Finally, it is worth noticing that in comprehensive derivations, conditions end up containing subformulas of A that do not logically follow from the instance α.

Definition 4.3. Let $\tau_1, ..., \tau_n$ be terms. We use the notation $\tau_1 - ... - \tau_n$, for $n \geqslant 1$, to indicate a list of separated terms, namely terms which cannot be united.

Definition 4.4. We define two sets of applications of explanatory rules:

- S_1 contains: rule $\circ_1$ applied on formulas of the form $C[A \wedge B]$, where $C[.] \in \mathcal{P}$, or on formulas of the form $C[A \vee B]$, where $C[.] \in \mathcal{N}$, as well as rule $\neg\circ_1$ applied on formulas of the form $C[\neg(A \vee B)]$, where $C[.] \in \mathcal{P}$ or on formulas of the form $C[\neg(A \wedge B)]$, where $C[.] \in \mathcal{N}$.
- S_2 contains: rule $\circ_1$ applied on formulas of the form $C[A \vee B]$, where $C[.] \in \mathcal{P}$ or on formulas of the form $C[A \wedge B]$, where $C[.] \in \mathcal{N}$, as well as rule $\neg\circ_1$ applied on formulas of the form $C[\neg(A \wedge B)]$, where $C[.] \in \mathcal{N}$ or on formulas of the form $C[\neg(A \vee B)]$, where $C[.] \in \mathcal{P}$.

To obtain sufficient reasons from comprehensive derivations, we employ the following labelling of derivations.

Definition 4.5. Let d be a comprehensive derivation of $\alpha \Rightarrow A$. The labelling of (the sequents in) d by lists of separated terms is defined inductively as follows:

1. For each leaf of d, if it is of one of the two following forms:

   (i) $X, \overbrace{\neg, ..., \neg}^{2n+1} p \mid \alpha \Rightarrow p$

   (ii) $X, \overbrace{\neg, ..., \neg}^{2n} p \mid \alpha \Rightarrow \neg p$

   for some integer n then associate with the leaf the empty set. Otherwise, associate with the leaf the set $\{p\}$ or $\{\neg p\}$, where p or $\neg p$, resepctively, are the literals on the right side of the sequent.

2. For any sequent in d, if one of the rules $\neg\neg$, $\circ_2, \neg\circ_2$ has been applied to obtain this sequent, and $\tau_1 - .... - \tau_n$ is the label associated with the premise of the rule, then $\tau_1 - .... - \tau_n$ is label associated with the sequent.

3. For any sequent in d, if any of the rules belonging to S_1 have been applied to obtain this sequent, and the labels associated to the premisses of the rule are $\tau_{n_1} - .... - \tau_{n_i}$ and $\tau_{m_1} - .... - \tau_{m_j}$, then the label associated to the sequent is the minimal sublist of $\tau_{n1} \cup \tau_{m1} - \tau_{n1} \cup \tau_{m2} - ... - \tau_{nk} \cup \tau_{ml}$.

4. For any sequent in d, if any of the rules belonging to S_2 have been applied to obtain this sequent, and the labels associated to the premisses of the node are $\tau_{n1} - .... - \tau_{nk}$ and $\tau_{m1} - .... - \tau_{mj}$, then the label associated to the sequent is the minimal sublist of $\tau_{n1} - .... - \tau_{nk} - \tau_{m1} - .... - \tau_{mj}$,

5. Consider the list of separated terms associated with the conclusion $\alpha \Rightarrow A$ of the derivation d. For any two sets τ_i and τ_j belonging to the list, it should never be the case that $\tau_i \subseteq \tau_j$. If so, eliminate the bigger set. Filtering this way the list, we obtain SR the set of all sets belonging to the (possibly filtered) list

where the minimal sublist of a list of separated terms $\tau_1 - .... - \tau_n$ contains all and only τ_i such that $\forall j \neq i,\ \tau_i \not\subseteq \tau_j$.

Definition 4.6. Let d be a comprehensive derivation of $\alpha \Rightarrow A$, equipped with the labelling defined above, and let $\tau_1 - .... - \tau_n$ be the label of the conclusion. $SR = \{\tau_i : i = 1, \ldots, n\}$.

The labelling in Definition 4.5 specifies inductively reasons for each sequent in the derivation. Since sufficient reasons have to be minimal, the definition works in such a way that all those reasons labelling the conclusion that are not minimal, i.e. such that they are subsumed by another reason labelling the conclusion, are filtered. Below we shall show that SR corresponds to the set of sufficient reasons for a decision $A(\alpha)$; it follows that the SR for $\alpha \Rightarrow A$ only depends on the sequent and is independent of the comprehensive derivation of this sequent used.

As an example, consider again classifier $\mathcal{C}$ represented by the formula $A = e \wedge (\neg f \vee w \vee g)$, and the case of Greg (the case of Susan can be treated analogously), which can be summed up by the following instance $\alpha = (e, \neg f, w, \neg g)$. Greg is admitted by classifier C: $\alpha \models A$. It is easy to check that the following is a comprehensive derivation d of $\alpha \Rightarrow A$:

$$\dfrac{\dfrac{\dfrac{e \wedge g \mid e, \neg f, w, \neg g \Rightarrow e \quad e \wedge g \mid e, \neg f, w, \neg g \Rightarrow \neg f}{e \wedge g \mid e, \neg f, w, \neg g \Rightarrow e \wedge \neg f}}{e, \neg f, w, \neg g \Rightarrow e \wedge (\neg f \vee g)} \qquad \dfrac{\dfrac{\dfrac{e \wedge g \mid e, \neg f, w, \neg g \Rightarrow e \quad e \wedge g \mid e, \neg f, w, \neg g \Rightarrow w}{e \wedge g \mid e, \neg f, w, \neg g \Rightarrow e \wedge w}}{e, \neg f, w, \neg g \Rightarrow e \wedge (w \vee g)}}{e, \neg f, w, \neg g \Rightarrow e \wedge ((\neg f \vee w) \vee g)}$$

Terms associated with d. The two top-left leafs are associated with the two lists of separated terms $\{e\}$ and $\{\neg f\}$, whilst the two top-right leafs are associated with the two lists of separated terms $\{e\}$ and $\{w\}$. Since for each pair of sequents associated with the lists of separated terms mentioned above a rule belonging to S_1 has been applied, we get on the one side $\{e, \neg f\}$ and on the other side $\{e, w\}$. In each case, we then have the rule $\circ_2$ followed by a rule belonging to S_2, hence we get the list of separated terms $\{e, \neg f\}$-$\{e, w\}$. Following the procedure of Definition 4.5, we conclude that $SR = \{\{e, \neg f\}, \{e, w\}\}$ is the set of two terms associated with derivation d of the sequent $e, \neg f, w, \neg g \Rightarrow e \wedge ((\neg f \vee w) \vee g)$.

As further illustration, consider now the classifier C', which is represented by the formula $A' = (p \wedge q) \vee (r \wedge \neg q)$, and which was mentioned in Section 2 as a more difficult example to treat. The sufficient reasons for instance $\alpha' = \{p, q, r\}$ are $\{p, q\}$ and $\{p, r\}$; we now show how our procedure is able to generate them. First of all, it is easy to check that the following is a comprehensive derivation d of $\alpha' \Rightarrow A'$:

$$\dfrac{\dfrac{\dfrac{\alpha' \Rightarrow p \quad \alpha' \Rightarrow r}{\alpha' \Rightarrow p \vee r} \quad \dfrac{\alpha' \Rightarrow q \quad \alpha' \Rightarrow r}{\alpha' \Rightarrow q \vee r}}{\alpha' \Rightarrow (p \wedge q) \vee r} \quad \dfrac{\dfrac{\neg q \mid \alpha' \Rightarrow p}{\alpha' \Rightarrow p \vee \neg q} \quad \dfrac{\neg q \mid \alpha' \Rightarrow q}{\alpha' \Rightarrow q \vee \neg q}}{\alpha' \Rightarrow (p \wedge q) \vee \neg q}}{\alpha' \Rightarrow (p \wedge q) \vee (r \wedge \neg q)}$$

Terms associated with d. The four top-left leafs are associated with the four lists of separated terms $\{p\}$, $\{r\}$, $\{q\}$ and $\{r\}$. Since for each pair of lists of separated terms, $\{p\}$ and $\{r\}$ on the one side, and $\{q\}$ and $\{r\}$ on the other side, a rule belonging to the set S_2 has been applied, we get $\{p\}$-$\{r\}$ and $\{q\}$-$\{r\}$, respectively. To the two sequents corresponding to these two lists of separated terms, a rule belonging to S_1 has been applied and thus we get, $\{p, q\}$-$\{p, r\}$-$\{r, q\}$- $\{r\}$. By filtering it, we get the minimal list $\{p, q\}$-$\{r\}$.

Let us now move to the two top-right leafs, which are associated with the two lists of separated terms $\{p\}$ and $\{\varnothing\}$. To each sequent corresponding to these lists the rule $\circ_2$ has been applied, so the two lists remain the same. Since, afterwards a rule belonging to the set S_1 has been applied, one gets $\{p\}$.

Now we have two sequents associated with the lists of separated terms $\{p, q\}$-$\{r\}$ and $\{p\}$, respectively. Since a rule belonging to the set S_1 has been applied on them, we get the terms $\{p, q\}$-$\{p, r\}$. Hence, following the procedure of Definition 4.5, we conclude that $SR = \{\{p, q\}$ and $\{p, r\}\}$ is the set of terms associated with the sequent $\alpha' \Rightarrow (p \wedge q) \vee (r \wedge \neg q)$.

5 Getting the complete reasons via the sequent calculus

In this section we show that, for any decision $A(\alpha)$, our procedure generates the set of all sufficient reason for it, or equivalently, its complete reason. To do so however, we first need to show some preliminary lemmas.

Lemma 5.1. *If SR, the set of terms associated to a comprehensive derivation of the sequent $X \mid \alpha \Rightarrow A$, only contains the term $\tau = \{\varnothing\}$, then the translation δ of $X \mid \alpha \Rightarrow A$, namely $(X \mid \alpha \Rightarrow A)^\delta := B_1^\perp \wedge ... \wedge B_n^\perp \rightarrow A$, is a tautology, where $X = \{B_1, ..., B_n\}$.*

Proof. Suppose d is a comprehensive derivation of the sequent $X \mid \alpha \Rightarrow A$ and SR, the set of terms associated with it, contains the empty set. We reason by induction on the height of d.

[-] Suppose $X \mid \alpha \Rightarrow A$ is an axiom, then, since the term associated with it is $\{\varnothing\}$, then $X \mid \alpha \Rightarrow A$ needs to be of either of one of the forms (i) or (ii) of Definition 4.5. It is straightforward to see that the translation δ of each of them is indeed a tautology.

[-] Suppose $X \mid \alpha \Rightarrow A$ is such that $A = C[B \wedge D]$ and $C[.] \in \mathcal{P}$ and the last rule of d is $\circ_1$. Then we have the following situation:

$$\frac{X \mid \alpha \Rightarrow C[B] \qquad X \mid \alpha \Rightarrow C[D]}{X \mid \alpha \Rightarrow C[B \wedge D]} \circ_1$$

where the lists of separated terms associated with each of the premises of this rule are $\{\varnothing\}$. By inductive hypothesis we have that $(X \mid \alpha \Rightarrow C[B])^\delta$ and $(X \mid \alpha \Rightarrow C[D])^\delta$ are tautologies. But then by logic, since $C[.] \in \mathcal{P}$, $(X \mid \alpha \Rightarrow C[B \wedge D])^\delta$ is a tautology.

[-] Suppose $X \mid \alpha \Rightarrow A$ is such that $A = C[B \vee D]$ and $C[.] \in \mathcal{P}$ and the last rule of d is $\circ_2$. Then we have the following situation:

$$\frac{X, C[D] \mid \alpha \Rightarrow C[B]}{X \mid \alpha \Rightarrow C[B \vee D]} \circ_2$$

where the list of separated terms associated with the premise is $\{\varnothing\}$. By inductive hypothesis we have that $(X, C[D] \mid \alpha \Rightarrow C[B])^\delta$ is a tautology. This means that $\bigwedge X^\perp \wedge C[D]^\perp \wedge \bigwedge \alpha \rightarrow C[B]$ is a tautology. But $\bigwedge X^\perp \wedge C[D]^\perp \wedge \bigwedge \alpha \rightarrow C[B]$ is equivalent to $\bigwedge X^\perp \wedge \bigwedge \alpha \rightarrow (C[D]^\perp \rightarrow C[B])$, which is in its turn equivalent to $\bigwedge X^\perp \wedge \bigwedge \alpha \rightarrow (C[D] \vee C[B])$, which is finally equivalent to $\bigwedge X^\perp \wedge \bigwedge \alpha \rightarrow C[B \vee D] = (X \mid \alpha \Rightarrow C[B \vee D])^\delta$, which is thus also a tautology.

[-] The cases of other rules are treated analogously. □

Lemma 5.2. *Consider the provable sequent $\alpha \Rightarrow A$. If the set SR associated to any comprehensive derivation of this sequent is $SR = \{\{\varnothing\}\}$, then A is a tautology.*

Proof. Straightforward from Lemma 5.1.

□

Lemma 5.3. *Consider the provable sequent $\alpha \Rightarrow A$. If A is of the form $B \vee \neg B$ or $\neg(B \wedge \neg B)$, for any B, then the set SR associated with any comprehensive derivation of this sequent only contains the term $\tau = \{\varnothing\}$.*

Proof. By induction on the complexity of A. We only consider the case where A is of the form $B \vee \neg B$, the other case being treated analogously. Note that having assumed that the sequent $\alpha \Rightarrow B \vee \neg B$ is provable, α either contains the literals of B or those of $\neg B$. In the proof, and for any analogous situation, we assume, without loss of generality, that α contains the literals of B.

[-] Suppose A is of the form $l \vee \neg l$. Then we have the following comprehensive derivation:

$$\dfrac{\dfrac{\neg l \mid \alpha \Rightarrow l'}{\vdots}}{\dfrac{\neg l \mid \alpha \Rightarrow l}{\alpha \Rightarrow l \vee \neg l}}$$

where the dots denote possible applications of the rule $\neg\neg$. As the axiomatic leaf of the comprehensive derivation above is of the form (i) of Definition 4.5, $SR = \{\{\varnothing\}\}$.

[-] Suppose A is of the form $C[l \wedge \neg l] \vee \neg C[l \wedge \neg l]$, where $l \wedge \neg l$ is the subformula of A with the highest depth which satisfies condition 2(a) of Definition 4.2. Then we have the following comprehensive derivation

$$\dfrac{\dfrac{\overset{i.h.}{\alpha \Rightarrow C[l] \vee \neg C[l]} \quad \alpha \Rightarrow C[l] \vee \neg C[l'] \; \vdots}{\alpha \Rightarrow C[l] \vee \neg C[l \wedge l']} \quad \dfrac{\overset{i.h.}{\alpha \Rightarrow C[l] \vee \neg C[l]} \quad \alpha \Rightarrow C[l] \vee \neg C[l'] \; \vdots}{\alpha \Rightarrow C[l'] \vee \neg C[l \wedge l']}}{\alpha \Rightarrow C[l \wedge l'] \vee \neg C[l \wedge l']}$$

Consider first the top-left branch of this derivation. By i.h. on the complexity of A, we have that the set τ associated with it is $\tau = \{\varnothing\}$. Since the rule operating on the conclusion of this branch belongs to S_2 (see Definitions 4.4, 4.5), the set associated to the conclusion of the rule is $\{\varnothing\}$ (as whatever list of terms is associated with the other premise, any such term is bigger than the emptyset). For a similar reasoning, the set associated with the second premise of the final rule is $\{\varnothing\}$. Since the last rule applied in the derivation above belongs to the set S_1, we obtain that the set $SR = \{\varnothing\}$, as desired. The other cases can be treated analogously.

□

Lemma 5.4. *Consider a sequent $X \mid \alpha \Rightarrow A$. Then, if a term τ belongs to the set SR associated with a comprehensive derivation of it, then $\tau \subseteq \alpha$ and $X^{\perp}, \tau \models A$.*

Proof. By induction on the comprehensive derivation of the sequent $X \mid \alpha \Rightarrow A$.

[-] Suppose $X \mid \alpha \Rightarrow A$ is a leaf, then it could be of either one of the forms (i) and (ii) of Definition 4.5, or in the axiomatic form $X \mid \alpha \Rightarrow \mathbf{p}$, where $\mathbf{p}$ is a literal. In the former case, $\tau = \{\varnothing\}$, and thus clearly $\tau \subseteq \alpha$. Moreover, straightforwardly, $X^{\perp}, \tau \models A$. In the latter case, $\mathbf{p} \in \alpha$, so $\tau = \{\mathbf{p}\} \subseteq \alpha$. Moreover, $X^{\perp}, \mathbf{p} \models \mathbf{p}$.

[-] Suppose $X \mid \alpha \Rightarrow A$ has been obtained by an explanatory rule, then we distinguish several cases. Suppose $A = C[B \wedge D]$ with $C[.] \in \mathcal{P}$ and it has been obtained by the rule $\circ_1$, then we have:

$$\dfrac{X \mid \alpha \Rightarrow C[B] \quad X \mid \alpha \Rightarrow C[D]}{X \mid \alpha \Rightarrow C[B \wedge D]}$$

By inductive hypothesis we have that for any τ' associated with the sequent $X \mid \alpha \Rightarrow C[B]$, $\tau' \subseteq \alpha$ and $X^{\perp}, \tau' \models C[B]$, but also that for any τ'' associated with the sequent $X \mid \alpha \Rightarrow C[D]$, $\tau'' \subseteq \alpha$ and $X^{\perp}, \tau'' \models C[D]$. By the way one construct lists of separated terms associated with sequents, we have that any τ associated with the sequent $X \mid \alpha \Rightarrow C[B \wedge D]$ is such that $\tau = \tau' \cup \tau''$ for some such τ' and τ''. Moreover, we have from the previous conditions and logic that $\tau \subseteq \alpha$ and $X^{\perp}, \tau \models C[B \wedge C]$.

Suppose $A = C[B \vee D]$ with $C[.] \in \mathcal{P}$ and it has been obtained by the rule $\circ_2$, then we have:

$$\dfrac{X, C[B] \mid \alpha \Rightarrow C[D]}{X \mid \alpha \Rightarrow C[B \vee D]}$$

By inductive hypothesis we have that for any τ' associated with the sequent $X, C[B] \mid \alpha \Rightarrow C[D]$, $\tau' \subseteq \alpha$ and $X^{\perp}, C[B]^{\perp} \tau' \models C[D]$. By the way one constructs lists of separated terms associated with sequents, we have that any τ associated with the sequent $X \mid \alpha \Rightarrow C[B \vee D]$ is such that $\tau = \tau'$ for some such τ'. Moreover, we have from the previous conditions and logic that $\tau \subseteq \alpha$ and $X^{\perp}, \tau \models C[B \vee D]$.The cases of the other forms of A and other rules are treated analogously. □

Theorem 5.5. *Consider the provable sequent $\alpha \Rightarrow A_\alpha$, as well as the set SR associated with a comprehensive derivation of this sequent. Any term τ belonging to SR is a sufficient reason for the decision $A(\alpha)$.*

Proof. Straightforward from Lemma 5.4 and the way the set SR is constructed. □

Lemma 5.6. *Consider the following decisions:*

for any $C[.] \in \mathcal{P}$	*(i)* $C[B \wedge D](\alpha)$	*(i)′* $C[B](\alpha)$, $C[D](\alpha)$
	(ii) $C[\neg(B \vee D)](\alpha)$	*(ii)′* $C[B^\perp](\alpha)$, $C[D^\perp](\alpha)$
for any $C[.] \in \mathcal{N}$	*(iii)* $C[B \vee D](\alpha)$	*(iii)′* $C[B](\alpha)$, $C[D](\alpha)$
	(iv) $C[\neg(B \wedge D)](\alpha)$	*(iv)′* $C[B^\perp](\alpha)$, $C[D^\perp](\alpha)$

where B and D are extended literals or formulas of the form $\overbrace{\neg\ldots\neg}^{m}(F \vee \neg F)$ or $\overbrace{\neg\ldots\neg}^{m}(F \wedge \neg F)$. If τ is a sufficient reason for any of the decisions on the left, then there exist two terms τ' and τ'', such that each is a sufficient reason for one of the corresponding decisions on the right and $\tau' \cup \tau'' = \tau$.

Proof. We prove the result for (i) and (i)′; the other cases are treated analogously. Suppose τ is a sufficient reason for $C[B \wedge D]_\alpha$; hence $\tau \models C[B \wedge D]_\alpha$ and so, by logic and the fact that B and D are extended literals or formula of the form $\overbrace{\neg\ldots\neg}^{m}(F \vee \neg F)$ or $\overbrace{\neg\ldots\neg}^{m}(F \wedge \neg F)$, $\tau \models C[B]_\alpha$ and $\tau \models C[D]_\alpha$. It follows that there are two sets τ' and τ'' such that $\tau' \subseteq \tau$ and $\tau'' \subseteq \tau$ and each is a sufficient reason for $C[B]_\alpha$ and $C[D]_\alpha$, respectively. Suppose $\tau' \cup \tau'' = \tau'''$ and $\tau''' \subseteq \tau$ (clearly τ''' cannot be bigger than τ for the way it has been constructed). Then there is a term τ''' such that $\tau''' \models C[B \wedge D]_\alpha$, $\tau''' \subseteq \alpha$ and τ''' subsumes τ. However, this contradicts the fact that τ is a sufficient reason for $C[B \wedge D]_\alpha$. Hence $\tau' \cup \tau'' = \tau$. □

Lemma 5.7. *Consider the following decisions:*

for any $B[.] \in \mathcal{P}$	*(i)* $B[C \vee D]_\alpha$	*(i)′* $B[C]_\alpha$, $B[D]_\alpha$
	(ii) $B[\neg(C \wedge D)](\alpha)$	*(ii)′* $B[C^\perp](\alpha)$, $B[D^\perp](\alpha)$
for any $B[.] \in \mathcal{N}$	*(iii)* $B[C \wedge D](\alpha)$	*(iii)′* $B[C](\alpha)$, $B[D](\alpha)$
	(iv) $B[\neg(C \vee D)](\alpha)$	*(iv)′* $B[C^\perp](\alpha)$, $B[D^\perp](\alpha)$

where C and D are extended literals or formulas of the form $\overbrace{\neg\ldots\neg}^{m}(F \vee \neg F)$ or $\overbrace{\neg\ldots\neg}^{m}(F \wedge \neg F)$. If τ is a sufficient reason for any of the decisions on the left, then it is also a sufficient reason for either both decisions on the right, or for just one of them.

Proof. The proof is analogous to that of the previous Lemma.

□

Theorem 5.8. *Let $A(\alpha)$ be a decision and let τ be a sufficient reason for it. Then τ belongs to the set SR associated with a comprehensive derivation of the sequent $\alpha \Rightarrow A_\alpha$.*

Proof. Let $A(\alpha)$ be a decision and let τ be a sufficient reason for it. We show that one can construct a comprehensive derivation of the sequent $\alpha \Rightarrow A_\alpha$ such that τ belongs to the set SR associated with it. We reason by main induction on the size of the set τ, and by secondary induction on the complexity of the formula A.

[-] Suppose $\tau = \{\varnothing\}$. Then, A_α is a tautology. We distinguish cases according to the form of A_α. If A_α is of the form $\overbrace{\neg \ldots \neg}^{m}(F \vee \neg F)$, where m is even, or $\overbrace{\neg \ldots \neg}^{n}(F \wedge \neg F)$, where n is odd, then, by potential applications of the $\neg\neg$ rule and Lemma 5.3, the set SR associated with the sequent $\alpha \Rightarrow A_\alpha$ only contains the empty set. Hence τ belongs to SR. If A_α is not of the form $F \vee \neg F$, then it must fit into one of the cases studied in the $\tau \neq \varnothing$ case below. For each case, the induction hypothesis can be applied to yield the conclusion that τ belongs to SR.

[-] Suppose $\tau \neq \varnothing$. We distinguish cases according to the form of A_α. Suppose A_α is an extended literal. Then the sequent $\alpha \Rightarrow A_\alpha$ is derivable from an axiom $\alpha \Rightarrow \overline{A_\alpha}$, where $\overline{A_\alpha}$ is a literal, perhaps with applications of the $\neg\neg$ rule. Thus, according to our procedure, we directly get the set $\tau = \{A_\alpha\}$, which contains all sufficient reasons for A_α and belongs to SR.

Suppose $A_\alpha = C[B \wedge D]$, $C[.] \in \mathcal{P}$, $B \wedge D$ is not of the form $E \wedge \neg E$, and B and D are extended literals or formulas of the form $\overbrace{\neg \ldots \neg}^{m}(F \vee \neg F)$ or $\overbrace{\neg \ldots \neg}^{m}(F \wedge \neg F)$. By Lemma 5.6, there exists two terms τ' and τ'' such that each of them is a sufficient reason for $C[B]$ and $C[D]$, respectively, and $\tau' \cup \tau'' = \tau$. By the inductive hypothesis, noting that τ', τ'' are not larger than τ and $C[B]$, C[D] are less complex than $C[B \wedge D]$, there exists a comprehensive derivation d_1 of $\alpha \Rightarrow C[B]_\alpha$ one of whose associated terms is τ', which belongs to the SR of $\alpha \Rightarrow C[B]_\alpha$, and there exists a comprehensive derivation d_2 of $\alpha \Rightarrow C[D]_\alpha$ one of whose associated terms is τ'', which belongs to the SR of $\alpha \Rightarrow C[D]_\alpha$. Applying the rule $\circ_1$ on the end-sequents of d_1 and d_2, we obtain a derivation d of the sequent $\alpha \Rightarrow C[B \wedge D]_\alpha$. d is a derivation of $\alpha \Rightarrow C[B \wedge D]_\alpha$, which is comprehensive by construction. By our procedure for associating terms to sequents, $\tau = \tau' \cup \tau''$ is among the terms associated to d; hence τ belongs to the SR associated with the sequent $\alpha \Rightarrow A_\alpha$.

Suppose $A_\alpha = C[B \vee D]$, $C[.] \in \mathcal{P}$, $B \vee D$ is not of the form $E \vee \neg E$, and B and D are extended literals or formulas of the form $\overbrace{\neg \ldots \neg}^{m}(F \vee \neg F)$ or $\overbrace{\neg \ldots \neg}^{m}(F \wedge \neg F)$. By Lemma 5.7, we know that τ, one of the sufficient reasons for $C[B \vee D]$, is (i) either also a sufficient reason for $C[B]$ and $C[D]$, or (ii) just for one of them. We show for (i) that we obtain the desired result; case (ii) is treated analogously. From (i), by inductive hypothesis, noting that τ', τ'' are not larger than τ and $C[B]$, C[D] are less complex than $C[B \wedge D]$, one can construct a comprehensive derivation d_1 of the sequent $\alpha \Rightarrow C[B]_\alpha$, one of whose associated terms is τ, which belongs to the SR of $\alpha \Rightarrow C[B]_\alpha$, and one can construct a comprehensive derivation of d_2 of $\alpha \Rightarrow C[D]_\alpha$ one of whose associated terms is τ, which belongs to the SR of $\alpha \Rightarrow C[D]_\alpha$. Applying the rule $\circ_1$ on the end-sequents of d_1 and d_2, we obtain a derivation d of the sequent $\alpha \Rightarrow C[B \vee D]_\alpha$, which by construction is comprehensive. By our procedure for associating terms to sequents, $\tau = \tau' \cup \tau''$ is among the terms associated to d; hence τ belongs to the SR associated with the sequent $\alpha \Rightarrow A_\alpha$. The cases of the other forms of A and other rules can be treated analogously. □

For any decision $A(\alpha)$, Theorems 3.8 and 3.9 guarantee that the explanatory calculus is able to decide whether the decision is positive or negative. If it is positive, a comprehensive derivation can be constructed (using a version of standard proof search: see Appendix A),

which, by Theorem 5.8, will provide the complete reason for it, i.e. the set of all sufficient reasons. In other words, the explanatory proof approach, applied to the machine learning classification problem, provides, in one stroke, a decision procedure for any decision $A(\alpha)$ as well as the complete reasons for it. Note that two elements of the explanatory calculus are crucial for this result. One is the possibility of working deep in formulas, the other is the presence of conditions beside reasons. The former allows decomposition of a formula along its inner parts, hence uncovering potential analytic subformulas that do not contribute to the collection of the compete reasons (such as the excluded middle in the example of classifier $\mathcal{C}'$). The latter allows one to separate those characteristics that are true for the instance under consideration from those that are not, i.e. those features that the applicant enjoys from those that she does not. It is thanks to these two elements that explanatory rules provide a natural tool to extract the sufficient reasons for a decision. Indeed it is easy to see that without these two elements the complete reasons for a decision can no longer be extracted: for instance, this is the situation if one replaces explanatory rules by standard inferential rules (or by other logics of explanations that have been recently proposed but that do not enjoy these features, e.g. see Fine (2012)). Both these elements were introduced with purely philosophical motivations: their technical value provides independent confirmation of their adequateness and strength. Finally, the elegant naturalness by means of which they apply in the framework of the theory developed by Darwiche and Hirth (2023) can be seen as the countersign of the deep relation between the two approaches.

6 Conclusions

In this paper we have investigated the links between two theories of explanation: one developed around the notion of prime implicant, the other, with illustrious philosophical roots, developed with the resources of proof theory. Not only is this interaction technically compelling, in that it establishes in a novel and original way a result that is neither immediate nor trivial, e.g. see Coudert and Madre (1993) and Shih et al. (2018). Moreover, and perhaps most importantly too, it seems to have a strong conceptual import as well. Indeed, it establishes that two approaches from completely different backgrounds based on different principles and perspectives coincide on what counts as sufficient reasons in the case of classification decisions. This is a priori so surprising that it suggests that there is something deep connecting them. This deep interaction is not only valuable per se but it could also give rise to the most fruitful and interesting paths of research, as results of each theory can be imported in the other. Indeed, if on the one hand the notion of implicant and prima implicant could be used to develop a semantics for the purely syntactic model developed by Poggiolesi (2024a), on the other hand, the resources of first-order logic, mobilized in Poggiolesi (2024a), may be profitably used to extend the theory of Darwiche and Hirth (2023).

References

Aristotle. *Posterior Analytics*. Oxford University Press, Oxford, 1993.

F. Berto and D. Nolan. Hyperintensionality. In E. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, pages 1–45. Stanford, 2023.

A. Betti. Explanation in metaphysics and Bolzano's theory of ground and consequence. *Logique et analyse*, 211:281–316, 2010.

K. Brünnler. *Deep inference and symmetry in classical proofs*. Logoc Verlag, 2004.

O. Coudert and J.-C. Madre. Fault tree analysis: 1020 prime implicants and beyond. In *Proceeding of the Annual Reliability and Maintainability Symposium*, pages 1–19. IEEE, 1993.

A. Darwiche. Three modern roles for logic in ai. In J. Lang, editor, *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 5103–5111. AAAI Press, 2018.

A. Darwiche and A. Hirth. On the (complete) reasons behind decisions. *Journal of Logic Language and Information*, 32:63–88., 2023.

M. Detlefsen. Fregean hierarchies and mathematical explanation. *International Studies in the Philosophy of Science*, 3:97–116., 1988.

K. Fine. Guide to ground. In F. Correia and B. Schnieder, editors, *Metaphysical grounding*, pages 37–80. Cambridge University Press, Cambridge, 2012.

F. Genco. Formal explanations as logical derivations. *Journal of Applied Non-Classical Logics*, 31:279–342, 2021.

A. Guglielmi and P. Bruscoli. On the proof complexity of deep inference. *ACM Transactions on Computational Logic*, 14:1–34, 2009.

C. Hempel. The function of general laws in history. *Journal of Philosophy*, 39:35–48., 1942.

C. Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press, New York, 1965.

A. Ignatiev, N. Narodytska, N. Asher, and J. Marques-Silva. From contrastive to abductive explanations and back again. In M. Baldoni and S. Bandini, editors, *AI*IA, volume 12414 of Lecture Notes in Computer Science*, pages 335–355. Springer, 2020.

H. Leitgeb. Hype: A system of hyperintensional logic. *Journal of Philosophical Logic*, 48: 305–405., 2019.

A. Liu and E. Lorini. A logic of *Black Box* classifier systems. In A. Ciabattoni, E. Pimentel, and R.J.G.B. de Queiroz, editors, *WoLLIC 2022*, pages 158–174. Lecture Notes in Computer Science, 2022.

X. Liu and E. Lorini. A unified logical framework for explanations in classifier systems. *Journal of Logic and Computation*, 33:485–515., 2023.

Paolo Mancosu, Francesca Poggiolesi, and Christopher Pincock. Mathematical explanation. In *Stanford Encyclopedia of Philosophy*, pages 1–43. Stanford, 2023.

G. Pimentel, R. Ramayanake, and B. Lellmann. Sequentialising nested systems. In M. Fitting, editor, *Tableaux 2019*, pages 147–165. LNCS 11714, 2019.

F. Poggiolesi. On defining the notion of complete and immediate formal grounding. *Synthese*, 193:3147–3167, 2016a.

F. Poggiolesi. A critical overview of the most recent logics of grounding. In F. Boccuni and A. Sereni, editors, *Objectivity, Realism and Proof*, pages 291–309. Boston Studies in the Philosophy and History of Science, Springer, Dordrecht, 2016b.

F. Poggiolesi. On constructing a logic for the notion of complete and immediate formal grounding. *Synthese*, 195:1231–1254, 2018.

F. Poggiolesi. Grounding and propositional identity: a solution to Wilhelm's inconsistencies. *Logic and logical philosophy*, 32:33–38, 2022.

F. Poggiolesi. (Conceptual) explanations in logic. *Journal of Logic and Computation*, pages 1–34, 2024a.

F. Poggiolesi and F. Genco. Conceptual (and hence mathematical) explanations, conceptual grounding and proof. *Erkenntnis*, 88:1481–1507, 2023.

Francesca Poggiolesi. Mathematical explanations: An analysis via formal proofs and conceptual complexity. *Philosophia Mathematica*, 32:1–30, 2024b.

J. Schaffer. Grounding in the image of causation. *Philosophical studies*, 173:49–100, 2016.

A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In J. Lang, editor, *IJCAI*, pages 5103–5111. AAAI Press, 2018.

A. Shih, A. Choi, and A. Darwiche. On the reasons behind decisions. In J. Lang, editor, *ECAI 2020*, pages 712–721. IOS Press, 2020.

A. S. Troelstra and H. Schwichtenberg. *Basic Proof Theory*. Cambridge University Press, Cambridge, 1996.

A Appendix

A.1 Algorithm to construct comprehensive derivation

The following algorithm returns a comprehensive derivation for a sequence $\alpha \Rightarrow A$ if it is derivable.

1. Take any subformula D of A of highest depth such that $D = D_1 \circ'' D_2$ and $R\circ'(D_1 \circ'' D_2)$ is a subformula of A satisfying the conditions in point 2(a) of Definition 4.2. Apply the appropriate $\circ_1$ or $\neg\circ_1$ rule with main connective $\circ''$.
2. **Repeat** while there exists D as specified in the previous step.
3. Take any B satisfying the conditions in point 3 of Definition 4.2, and apply the appropriate $\circ_1$ or $\neg\circ_1$ rule with any of these as the main connective.
4. **Repeat** while there exists such B.
5. Take any leaf of the form $\alpha \Rightarrow l$ for an extended literal l that is not a literal, and apply the $\neg\neg$ rule.
6. **Repeat** while there exist such leaves.

   At this point, the algorithm has constructed a tree with, as each leaf, a formula of the form $\alpha \Rightarrow l$ for some literal l.

7. Take any leaf labelled by $\alpha \Rightarrow l$ with $l \notin \alpha$.

(a) **If** the first $\circ_1$ or $\neg\circ_1$ rule applied below this leaf does not satisfy the conditions for the application of a $\circ_2$ or $\neg\circ_2$ rule (Definition 3.6), then go down the tree to the first application of an $\circ_1$ or $\neg\circ_1$ rule satisfying the conditions in Definition 3.6 below this leaf, and replace it with the corresponding $\circ_2$ or $\neg\circ_2$ rule respectively, where the formula derived from the leaf enters the condition.

(b) **If** the first $\circ_1$ or $\neg\circ_1$ rule applied below this leaf does satisfy the conditions for the application of a $\circ_2$ or $\neg\circ_2$ rule (Definition 3.6), then go down the tree to the last consecutive application of this rule and replace it with the corresponding $\circ_2$ or $\neg\circ_2$ rule respectively, where the extended literal from the leaf enters the condition. Re-apply the initial $\circ_1$ or $\neg\circ_1$ rule and step 5 to obtain a tree whose leaves are literals.

8. **Repeat** while there exist such leaves.

9. **If** there is a leaf labelled by $\alpha \Rightarrow l$ with $l \notin \alpha$, then **Return**: $\alpha \Rightarrow A$ is not valid.

10. **If** not, then **Return** the derivation (tree).

If case 10 is satisfied, the algorithm returns a well-formed derivation (i.e. a derivation starting with axioms, where every rule is applied correctly). By construction, it satisfies the conditions in Definition 4.2, i.e. it is a comprehensive derivation.

Since this is a standard proof search algorithm adapted for the explanatory calculus **EC**, it not only produces a comprehensive derivation where it exists, but also identifies whether the sequent $\alpha \Rightarrow A$ is valid or not, i.e. whether α is an instance of A. Its complexity is comparable to standard proof search procedures, e.g. see Troelstra and Schwichtenberg (1996).