(Conceptual) Explanations in logic

Francesca Poggiolesi

Francesca Poggiolesi,
CNRS, IHPST UMR 8590

(Conceptual) explanations in logic

Abstract

To explain phenomena in the world is a central human activity and one of the main goals of rational inquiry. There are several types of explanation: one can explain by drawing an analogy, as one can explain by dwelling on the causes (see e.g. see Woodward (2004)). Amongst these different kinds of explanation, in the last decade philosophers have become receptive to those explanations which explain by providing *the reasons* (or the grounds) *why* a statement is true; these explanations are often called *conceptual explanations* (e.g. see Betti (2010)). The main aim of the paper is to propose a logical account of conceptual explanations. We will do so by using the resources of proof theory, in particular sequent rules analogous to *deep inferences* (e.g. see Brünnler (2004)). The results we provide not only shed light on conceptual explanations themselves, but also on the role that logic and logical tools might play in the burgeoning field of inquiry concerning explanations. Indeed, we conclude the paper by underlining interesting links between the present research and some other existing works on explanations and logic that have arise in recent years, e.g. see Arieli et al. (2022); Darwiche and Hirth (2023); Piazza et al. (2023).

> For here it is for the empirical scientist to know the fact and for the mathematical to know the *reason why*.[1] Aristotle (1993).

1 Introduction

To explain phenomena in the world is a characteristically human enterprise and a central goal of rational inquiry; it is thus no surprise that the notion of explanation has been one of the most intensely discussed topics in philosophy of science over the past century, and computer science over the past decade. Though the term covers a wide range of diverse cases – from explaining how to build an Ikea bookcase, or what a symbol means, to explaining a new concept to a child – doubtless the main ones for human endeavor are explanations that have a deductive form and aim to shed light on why a certain phenomenon occurs or why a certain proposition is true. Archetypal examples are *causal explanations* (Woodward, 2004; Pearl, 2000), which explain their target phenomena by providing their causes. However, it has recently become increasingly clear that many compelling examples of deductive explanations-why cannot be captured by causal accounts. In physics as well as in mathematics, several types of explanations arise that do not seem to rely on any causal mechanism: very naturally, the idea that causation, though certainly a key ingredient of explanation, is probably not the full story, started to spread; non-causal explanations, namely explanations that in one way or another go beyond causation, have become a new thrilling and thriving subject of research.[2]

[2] E.g. see Lange (2017).

Figure 1: Toy examples of conceptual explanations

Informal	Example 1.1.	Example 1.2.	Example 1.3.
Formal	$\neg(p \vee q)$	V	$\forall x(SC(x) \rightarrow IC(x))$
	because	because	because
	$\neg p$ and $\neg q$	F_1 and F_2	$\forall x(SC(x) \rightarrow C(x))$ and $\forall x(SC(x) \rightarrow T(x))$

p:= "it is raining," q:= "it is windy,"V:= "this animal is a vixen," F_1:= "this animal is a female," F_2:= "this animal is a fox," $SC(x)$:= "x is a Stanford computer science graduate," $IC(x)$:= "x is an ideal candidate for a certain tech job," $C(x)$:= "x has coding competences," $T(x)$:= "x has team-working skills."

Here we focus on a type of non-causal explanation that has been receiving increasing attention in the philosophical literature: *conceptual explanations* (e.g. see Betti (2010); Detlefsen (1988); Hunmean (2010); Mancosu et al. (2023)). Recognition of such explanations dates back millennia; as illustrated by the epigraph, a long tradition of scholars, including Aristotle, Proclus, Leibniz, Arnauld and Nicole, Bolzano, Frege,[3] have argued for their importance for the scientific inquiry. Conceptual explanations bear a striking analogy to causal explanations: where the latter seeks to explain by providing the causes, the former explains why a proposition is true by identifying the reasons, or the grounds,[4] for its truth. Instances of conceptual explanations range from stock, toy examples to more involved, real-life ones. The argument which explains why (1.1.) it is not the case that it is raining or it is windy because it is neither raining nor it is windy (together with the definition of the disjunction) is an example of conceptual explanation. Indeed it displays the reasons, rather than the causes, for the conclusion, with the relation between reasons and conclusion holding in virtue of the concepts - mainly *or* - that they contain. Similarly, the explanation of why (1.2.) a certain animal is a vixen that evokes that animal being a female as well as it being a fox (together with the definition of vixen), is an example of conceptual explanation. Indeed it displays the reasons, rather than the causes, of why that animal is a vixen, with the relation between reasons and conclusion holding in virtue of the concepts - *vixen*, *female* and *fox* - that they contain. Finally, the explanation why (1.3.) all Stanford Computer Science graduates are ideal candidates for a particular tech job in terms of their coding competences and teamwork skills (together with the stipulation of what an ideal candidate for that position is) is another example of a conceptual explanation in that it explains why a certain conclusion is true by bringing out the reason(s) for its truth.

Beyond simple examples, there are many instances of conceptual explanation with more refined (logical) structure, in particular involving quantifiers. This is especially the case for *mathematical explanations* – and more precisely those proofs in mathematics that not only show a theorem to be true, but also seem to provide the reason(s) or the ground(s) why it is true. As frequently noted, this kind of mathematical explanations could be argued to count among conceptual explanations (Betti, 2010; Mancosu et al., 2023). A simple example is the (elementary) proof which explains why (1.4) zero or the successor of any natural number is a natural number by emphasizing that zero is a natural number, and any successor of a

[3]E.g. see Detlefsen (1988).

[4]In this paper, we take the words "reason" and "ground" as synonymous: they are both taken to denote the objective foundations of truths. On this important point see the more extensive discussion later in this Section.

Figure 2: Mathematical examples of (conceptual) explanations

Informal	Example 1.4	Example 1.5.	Example 1.6.
Formal	$\forall x(Z(x) \vee SN(x) \rightarrow N(x))$	$\forall x\forall y\forall z\forall w((Circ(z,x,xy) \wedge Circ(w,y,xy) \wedge Point(x) \wedge Point(y)) \rightarrow \exists k\ (Point(k) \wedge k \in z \wedge k \in w \wedge l(kx) = l(ky) = l(xy)))$	$\forall x\forall y\forall z\forall w(RTr(xyz - xwz//xwy) \rightarrow xz^2 + xy^2 = zy^2)$
	because	because	because
	$\forall x(Z(x) \rightarrow N(x))$ and $\forall x(SN(x) \rightarrow N(x))$	$\forall x\forall y((Point(x) \wedge Point(y)) \rightarrow \exists k(\ Point(k) \wedge l(kx) = l(ky) = l(xy))$	$\forall x\forall y\forall z\forall w(Sim(xyz,xwz) \rightarrow xz^2 = zw.zy)$ and $\forall x\forall y\forall z\forall w(Sim(xyz,xwy) \rightarrow xy^2 = wy.zy)$

$Z(x)$:= "x is zero," $SN(x)$:= "x is the successor of a natural number," $N(x)$:= "x is a natural number," $Circ(x,z,zy)$:= "x is a circle, with center z and radius zy," $Point(x)$:= "x is a point," $l(xy)$:= "xy is the length between point x and point y," $RTr(xyz - xwz//xwy)$:= "xyz is a right angle triangle divided into two triangles xwz and xwy by the height," $Sim(xyz,xwz)$:= "the triangles xyz and xwz are similar."

natural number is a natural number. The literature on mathematical explanations is replete with less trivial examples (see e.g. Mancosu et al. (2023)). An old one dates back to Bolzano (2014), who analyzes the theorem which states that (1.5.) given any two circles A and B, one with center a and radius ab, and the other with center b and radius ab, then there always exists a point c where they intersect such that $l(ac) = l(cb) = l(ab)$. There exists a proof of this theorem that crucially relies on the property that, for any two points a and b, there always exists a third point c such that $l(ab) = l(bc) = l(ac)$. Bolzano argues that this proof is explanatory insofar as it relies on the relation between the property of the points - the reason - and the property of the circles - the conclusion. In its turn, this relation holds in virtue of the concepts involved, namely the concepts of point, radius, circle.

Or, consider an example from one of the groundbreaking articles in the literature on mathematical explanation (Steiner, 1973) concerning Pythagoras' theorem. Out of the many proofs of the theorem, one that is often argued to be explanatory crucially relies on a property of similar triangles. Following Steiner (1973), but also a more recent and detailed analysis put forward by Poggiolesi (2024), this proof is explanatory in that it relies on the relation between a certain property of similar triangles - the reason - and a certain property of right-angled triangles - the conclusion. In particular, the reason why any right-angle triangle ABC, divided in the two triangles ABH and ACH by the height, is such that the square of the hypothenuse AB is equal to the sum of the squares of the two sides AB and AC is that the similarity between ABC and each of the triangles ABH and ACH involves certain ratios amongst their sides. In turn, the relation between Pythagoras' theorem and its reason holds in virtue of the concepts that these elements contain, namely the concepts of similarity, and right-angled triangle. Hence, Pythagoras' mathematical explanatory proof is also a paradigmatic example of conceptual explanation, in that it display the features of this type of explanation.

Despite their widespread relevance, to date conceptual explanations have received little attention in logic. This absence of systematic study is all the more striking given the recent interest in logic in non-causal explanations, be it in the field of XAI (e.g. see Darwiche and Hirth, 2023; Shih et al., 2018), or for abductive reasoning (Arieli et al., 2022; Arieli and

Strasser, 2015; Piazza et al., 2023; Pulcini and Varzi, 2021). As a result, this is an important gap in the logical literature and the main goal of the present paper is to fill it. In particular our aim is to develop a logical theory of conceptual explanations, which is strong enough to encompass the several different cases of conceptual explanations.

There are (at least) two main questions that a logical theory of (conceptual) explanations need to address: (i) What is the structure of conceptual explanations? And (ii) what are the reasons for a certain conclusion? By relying on previous research in Poggiolesi (2016b, 2018), which is, as far as we know, the only one that have previously addressed analogous questions, we provide an answer to both (i) and (ii). As for (i) we provide a modelization of conceptual explanations in terms of proofs. In particular, we introduce explanatory rules which are such that not only the conclusion is inferable from the premisses, but also the premisses represent the reasons why the conclusion is true. The concatenation of these rules represent the logical structure behind conceptual explanations. Note that, as the examples above show (see Figures 1 and 2), reasons (or grounds) are often linked to their conclusion by operating inside formulas. Take for instance the example 1.4. (see Figure 2), where $\forall x((Zx \vee SNx) \rightarrow Nx)$ is explained by $\forall x(Zx \rightarrow Nx)$ and $\forall x(SNx \rightarrow Nx)$. The link between these formulas occurs deep inside the formulas themselves: in particular, the connective $\vee$ inside the explanandum is broken into two and thus give rise to $\forall x(Zx \rightarrow Nx)$ and $\forall x(SNx \rightarrow Nx)$. As a consequence, explanatory rules will have the form of *deep inferences*, namely a recently introduced variation of the sequent 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 for (ii), we will set out features which establish when some formulas can be seen as the reasons (or grounds) why another is true. In other terms, by relying on deep insights which can be found in the philosophical literature, we will set out those conditions that are necessary and sufficient to ensure that some formulas can be seen as the reasons, or the grounds, why another is true. Moreover, we will show that the answer to this question is fully coherent with the answer to question (i), by proving that our explanatory rules provide all and only those relations from reasons (or grounds) to conclusion that satisfies the proposed features.

Note that a logical theory of conceptual explanations, and related relation from reasons to conclusion (i.e. related grounding relation), will limit itself to the logical form of these objects. For instance, when considering examples 1.2, 1.3, 1.5 and 1.6 it is straightforward to note that they all have an analogous structure; it is this structure that the paper analyses. In other words, although the paper focuses on the realm of conceptual explanations and related grounding relation, it only deals with the formal part that characterize these objects. However, their explanatory power also relies on the relations amongst the concepts that they involve; the analysis of these relations goes beyond the scope of this paper and will thus be left aside.[5]

As already noted, the present work draws on the research developed in Poggiolesi (2016b, 2018), going beyond it both on the conceptual and the formal front. The cited papers contribute to a more established literature dedicated to grounding (and, amongst other issues, to its relationship with explanation), developed notably in metaphysics over the last two decades. More recently, however, there has been an increasing interest in the role and status of explanation – and secondarily grounding – in its relation to the science (e.g., see Betti (2010); Kortabarria and Giannotti (2024); Poggiolesi and Genco (2023)). Such a shift

[5]Details of such an analysis are provided by Poggiolesi and Genco (2023); Poggiolesi (2024).

is not unfamiliar in philosophy and logic; on the contrary, it is analogous to contrasts between metaphysical analysis of concepts such as causality or necessity (e.g. see for references Gallow (2021); Kment (2021)) and formal studies of these concepts (e.g., see Pearl (2000); Blackburn et al. (2001)). Such a shift naturally implies a focus on different questions and methods concerning the notions at issue: in our case, questions about the nature of grounding and metaphysical tools for analyzing them make way for a focus on issues such as the structure of explanations, for which conceptual analysis and formal tools are more relevant. As is clear from the questions stated above, the present paper sits squarely in this new branch of the literature, focusing on structural questions and using formal tools, whilst remaining non-committal on metaphysical considerations. In concordance with this focus, the word *reason*, rather than *ground*, which is more closely associated with the metaphysical literature, will be used in this paper.

Beyond a shift in research questions and methods, the focus on explanation and reasons in the sciences brings to the fore a new set of examples, as compared to the traditional grounding literature in metaphysics. Typically, this literature discusses almost exclusively toy cases such as "the ball is red and the ball is round because the ball is red and round" (e.g. see Fine (2012)) at any length. By contrast, bringing in explanations in fields such as mathematics – whereby a certain theorem, say, is explained by displaying the reasons why it is true – requires consideration of much more intricate instances of explanations and reasons. The ones which are presented in Figures 1 and 2 are illustrative of the kind of complexity that needs to be addressed.

As a result, not only does the present work differ from Poggiolesi's one in the main object of concern – namely explanations rather than grounding – as well as in the background working framework – sciences rather than metaphysics –, but crucially in the formal theory it proposes. More specifically, there are two main novelties that this work presents and are absent from Poggiolesi (2016b, 2018): firstly, the current paper works in first-order logic; secondly, and more importantly, it employs a formalization that works deep inside formulas. Moreover, whilst Poggiolesi's work relies on the resources of natural deduction calculi, results in this paper are based on the means of the sequent calculus, which is a notoriously stronger and more powerful proof-theoretic tool. Hence, summing all these features up, the present paper offers a very general theory of (conceptual) explanations which doesn't have any analogue in the literature. At the end of the paper, we discuss how this theory relates to existing work on explanation, and how it opens up interesting paths of future research.

The paper is organized in the following way. In *Section 2* we will formalize the notion of conceptual explanation via the relation of formal explanation, that is defined via explanatory rules, i.e. rules that provide the reasons why their conclusion is true and that are added to the classical sequent calculus for first-order logic. In *Section 3* we prove that explanatory rules are admissible in the classical sequent calculus, i.e. explanatory rules serve to construct derivations with an explanatory power, not to prove new theorems. *Section 4* will serve to set those features according to which some formulas count as the reasons why another is true, and in *Section 5* we will prove that the explanatory rules provide all, and only, relations from reasons to conclusion. Whilst in *Section 6* we will emphasize some interesting links with other related works on explanations, in *Section 7* we will draw some conclusions and sketch directions of future research.

2 Formal explanations

In order to provide the formal structure which underlies conceptual explanation, we start from an idea that is both ancient and central in the literature: explanations can be seen as deductive arguments which, starting from true premisses - be they the causes or the grounds - explain a certain conclusion.[6] Of course not any deductive argument constitutes an explanation, but some of them do, namely those which have an explanatory power. The perspective that we will develop here[7] 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 mathematical explanations, a notable subset of conceptual explanations, actually are proofs of mathematical theorems, which show why those theorems are true.

Let us pursue this perspective further. Since proofs are standardly formalized in logic by means of *derivations*, we will formalize conceptual explanations as a special type of derivations. More precisely, we will introduce a metalinguistic relation called *formal explanation*,[8] denoted by the symbol $\Vdash$, which will represent the formal counterpart of conceptual explanations, as well as a special case of the standard notion of derivation. As derivations are introduced via inferential rules, formal explanations will be introduced via explanatory rules, namely rules where not only the conclusion is inferable from their premise(s), but also such that the premisses are the (formal) reasons why the conclusion is true. We will consider explanatory rules, and related formal explanation relation, in first-order classical logic.

Definition 2.1. The language of first-order logic, $\mathcal{L}$, is composed by: variables (x_0, x_1, x_2, ...), constants (c_0, c_1, c_2, ...), predicates (P_0^k, P_1^k, P_2^k, ...), logical connectives ($\neg$, $\wedge$, $\vee$), quantifiers ($\forall$, $\exists$), and parentheses: (,). We take the symbols $\top, \bot$ and $\rightarrow$ to be defined as usual. For the sake of simplicity we do not use the identity symbol nor the functional symbols. Also we will use the symbols $\circ$ and $\odot$ in the following way: $\circ = \{\wedge, \vee\}$ and $\odot = \{\forall, \exists\}$. The set of well-defined formulas, $\mathcal{WF}$, is constructed in the standard way. A closed formula, or a sentence, is a formula where no free variable occurs. The set of closed formulas of $\mathcal{L}$ will be denoted by $\mathcal{CF}$.

Definition 2.2. Given, the multiset $M \subseteq \mathcal{WF}$ and formula $A \in \mathcal{WF}$, we use the standard notation, $M \models A$, to mean that A logically follows from M in first-order classical logic. The notation $M \vdash A$ means that there exists a derivation from M to A in (the standard sequent calculus for) first-order classical logic.

In order to properly spell out the notion of (conceptual) explanation under consideration, we introduce some notable distinctions that help identifying different types of deductive explanations. Here we start with the following two, namely the distinction between *total/partial* explanations, and the distinction between *immediate/mediate* explanations.[9] A total explanation is one that provides all the reasons why something is true. In other terms, the multiset of all, and only, those formulas each of which contributes to explain C is a total explanation of C. On the other hand, each proper sub-multiset of the total explanation of C is a partial explanation of C.[10]

[6] E.g. see Aristotle (1993); Hempel (1965, 1942).
[7] See also Poggiolesi and Genco (2023).
[8] This name has already been used in Poggiolesi (2018). Here we employ it in a broader sense.
[9] E.g. see Lewis (1973); Schaffer (2016).
[10] E.g. see also Poggiolesi (2020b).

Figure 3: The sequent calculus $\mathbf{Gcl}^+$.

$$p, M \Rightarrow N, p \qquad \frac{M \Rightarrow N}{P \Rightarrow Q \mid M \Rightarrow N}\ cw$$

$$\frac{M \Rightarrow N, F}{\neg F, M \Rightarrow N}\ \neg L \qquad \frac{F, M \Rightarrow N}{M \Rightarrow N, \neg F}\ \neg R \qquad \frac{F, G, M \Rightarrow N}{F \wedge G, M \Rightarrow N}\ \wedge L \qquad \frac{M \Rightarrow N, F \quad M \Rightarrow N, G}{M \Rightarrow N, F \wedge G}\ \wedge R$$

$$\frac{\forall x F, F(x/t), M \Rightarrow N}{\forall x F, M \Rightarrow N}\ \forall L \qquad \frac{M \Rightarrow N, \forall x F(x/y)}{M \Rightarrow N, \forall x F}\ \forall R$$

where in $\forall R$ y does not occur in M nor in N.

As concerns the other distinction, whilst an immediate explanation is one that involves a single explanatory step, i.e. a step that does not seem to be further reducible, a mediate explanation includes several consecutive immediate steps. In this paper we will first deal with the notion of total and immediate formal explanation and then generalize it to the mediate case.

There exists a third distinction that is linked to the notion of total explanation and that arises both in the causal and conceptual framework. To illustrate it, we start from the causal case, where it is most well-known. Consider the following notorious example.[11] Billy and Suzy throw rocks at a bottle, which shatters. A causal explanation of why the glass shattered is that Suzy threw her rock at it. Indeed since Suzy threw her rock first, her rock arrived first and shattered the glass; Billy's rock sailed past the already-broken bottle. Billy's throw is thus not a cause, but only a potential cause of why the bottle shattered. Potential causes are central for total explanations: if Billy's rock hit the bottle at the same time as Suzy's rock, it would have been part of the total explanation of why the glass shattered.

A distinction analogous to that between causes and potential causes also arises in the conceptual framework. Consider indeed the following situation. Billy is Jane's brother and Suzy is Jane's sister. Jane 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 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 a girl, his having a girl 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*.[12] 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 girl.

In order to define the notion of formal explanation, we work with the classical sequent calculus for first-order logic, implemented with the metalinguistic symbol "|", for conveying conditions, and the related rule *cw* which allows to introduce conditions beside standard sequents. Conditions only play a role in explanatory rules - no inferential rule operates on conditions - hence, the sequent calculus $\mathbf{Gcl}^+$ (see Figure 3) is equivalent to the standard classical sequent calculus for first-order logic **Gcl** (see Troelstra and Schwichtenberg (1996)). The notion of sequent, its interpretation, and the interpretation of inferential rules are

[11] E.g. see Menzies and Beebee (2020).

[12] Here we borrow this terminology from Genco (2021).

standard (e.g. see again Troelstra and Schwichtenberg (1996)). We call *c-sequent* a sequent that only contains closed formulas.

We will add to $\mathbf{Gcl}^+$ explanatory rules. As already underlined, explanatory rules provide the (total and immediate) reasons why their conclusion is true; but the link between reasons and conclusion often require looking deep inside formulas (see examples in Figures 1 and 2). So in order to be able to introduce explanatory rules, we will first need to introduce the notations necessary to work inside formulas. We will do so with the notions of *context* and *formula in a context*. Roughly speaking, a context is the part of a formula that one does not focus on, and is denoted with the notation C[.]. For example consider the first-order formula $F = \exists x(Sx \wedge Tx) \vee \forall x \forall y(Px \rightarrow Qx \wedge Ry)$ and suppose we want to focus on a particular part of F, say $Qx \wedge Ry$. We denote this fact by rewriting F as $C[Qx \wedge Ry]$, where $C[.]$ is the context $\exists x(Sx \wedge Tx) \vee \forall x \forall y(Px \rightarrow .)$ and $Qx \wedge Ry$ is the formula in the context $C[.]$. Note that when working in an explanatory framework, negation needs to be handled with particular care. 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 $\neg(\neg p \vee \neg q)$. As discussed in Poggiolesi (2016a, 2022), whilst the (total) reasons of $\neg(p \vee q)$ amount to the formulas $\neg p, \neg q$, the (total) 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 said. To avoid such undesirable cases, we define contexts as only involving 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*, introduced below.

Definition 2.3. The converse of a formula A, written A^*, is defined as follows:

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

where the main connective in E is not a negation, $n \geqslant 0$ and 0 is taken to be an even number. For any multiset M, $(M)^* := \{B^* : B \in M\}$.

So, for instance, returning to our previous example, the converse of $\neg p$, namely $(\neg p)^*$, is p, not $\neg\neg p$.

Definition 2.4. The set Co of contexts is inductively defined in the following way:

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

Definition 2.5. For all contexts $C[.]$, and formulas F, we define $C[F]$, a formula in a context, as follows:

- if $C[.] = [.]$, then $C[F] = F$,
- if $C[.] = \neg\neg D[.]$, then $C[F] = \neg\neg D[F]$,
- if $C[.] = D' \circ D[.]$, $D[.] \circ D'$, $\odot x D[.]$, $\neg D[.]$, then $C[F] = D' \circ D[F]$, $D[F] \circ D'$, $\odot x D[F]$, $\neg D[F]$, 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 2.6. 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.

(i) $[.] \in \mathcal{P}$,

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

(ii) $\neg G^-$, $F \wedge G^+$, $G^+ \wedge F$, $F \vee G^+$, $G^+ \vee F$, $\forall x G^+$, $\exists x G^+ \in \mathcal{P}$.

(iii) $\neg G^+$, $F \wedge G^-$, $G^- \wedge F$, $F \vee G^-$, $G^- \vee F$, $\forall x G^-$, $\exists x G^- \in \mathcal{N}$

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

The last ingredient needed to introduce explanatory rules is obtained by defining the scope of a context, in terms of the quantifiers that formulas in contexts lie in.

Definition 2.7. If $C[.]$ is a context, then the *scope of a context*, $SC(C)$, and the *inverse scope* $SC^{inv}(C)$, are defined inductively in the following way:

- if $C[.]= [.]$, then $SC(C) = SC^{inv}(C) = \varnothing$,
- if $C[.]= \neg\neg D[.]$, then $SC(C) = SC(D)$ and $SC^{inv}(C) = SC^{inv}(D)$,
- if $C[.]= D' \circ D[.]$ or $D[.] \circ D'$, then $SC(C) = SC(D)$ and $SC^{inv}(C) = SC^{inv}(D)$,
- if $C[.]= \forall x D[.]$, then $SC(C) = \forall x.(SC(D))$ and $SC^{inv}(C) = \exists x.(SC^{inv}(D))$,
- if $C[.]= \exists x D[.]$, then $SC(C) = \exists x.(SC(D))$ and $SC^{inv}(C) = \forall x.(SC^{inv}(D))$
- if $C[.]= \neg D[.]$, then $SC(C) = SC^{inv}(D)$ and $SC^{inv}(C) = SC(D)$.

Note that scopes of contexts are defined in such a way that an existential that occurs in a context with a negative polarity is transformed into an universal in a context with a positive polarity, whilst an universal in a context with a negative polarity is transformed into an existential in a context with a positive polarity. For instance, consider a context $C[.] = \forall x \neg \exists y \neg [.]$, then $SC(C)= \forall x \forall y$. On the other hand, consider a context $C[.] = \exists x \neg \forall y \neg [.]$, then $SC(C)= \exists x \exists y$.

Given a formula in a context $C[F]$, we can restrict the scope of context $SC(C)$ to the *scope of context relative to the formula* F, $SC_F(C)$, depending on the variables that the quantifiers in $SC(C)$ bound in F.

Definition 2.8. Let F be a formula of $\mathcal{L}$, then the free variables of F, $FV(F)$, are standardly defined as those variables occurring in F which are not bound by any quantifier. We define the *restricted free variables* of a formula F, $FV^+(F)$, in the following way

- if F is not of the form $G \circ G'$, then $FV^+(F) = FV(F)$,
- if F is of the form $G \circ G'$, then $FV^+(F) = FV(G) \cap FV(G')$.

Figure 4: Explanatory propositional rules.

$$\frac{M \Rightarrow N, C[F]}{M \Rightarrow N, C[\neg\neg F]}\ \neg\neg$$

$$\frac{M \Rightarrow N, C[F] \quad M \Rightarrow N, C[G]}{M \Rightarrow N, C[F \circ G]}\ \circ_1 \qquad \frac{M \Rightarrow N, C[F_j] \mid \quad M \Rightarrow N, C[F_i]}{M \Rightarrow N, C[F_1 \circ F_2]}\ \circ_2$$

$$\frac{M \Rightarrow N, C[F^*] \quad M \Rightarrow N, C[G^*]}{M \Rightarrow N, C[\neg(F \circ G)]}\ \neg\circ_1 \qquad \frac{M \Rightarrow N, C[F_j^*] \mid \quad M \Rightarrow N, C[F_i^*]}{M \Rightarrow N, C[\neg(F_1 \circ F_2)]}\ \neg\circ_2$$

where $i, j = \{1, 2\}$ and $j \neq i$.

Definition 2.9. Let $C[F]$ be a formula in a context, and let $FV^+(F) = x_1, ..., x_n$. The scope of a context $C[.]$ relative to the formula F, $SC_F(C)$, is the result of removing from $SC(C)$ any quantifier that is not of the form $\odot x_1, ..., \odot x_n$.

We can classify any formula in context $C[F]$ according to the polarity of (the formula in) the context and the type of quantifiers that its $SC_F(C)$ corresponds to.

Definition 2.10. For any formula in a context $C[F]$, we say that it has

a *positive universal scope* (PUS) if $C[.] \in \mathcal{P}$ and $SC_F(C) = \forall x_1, ...\forall x_n$,

a *negative universal scope* (NUS) if $C[.] \in \mathcal{N}$ and $SC_F(C) = \forall x_1, ...\forall x_n$,

a *positive existential scope* (PES) if $C[.] \in \mathcal{P}$ and $SC_F(C) = \exists x_1, ...\exists x_n$,

a *negative existential scope* (NES) if $C[.] \in \mathcal{N}$ and $SC_F(C) = \exists x_1, ...\exists x_n$.

We now have all the elements to introduce explanatory rules. In Figure 4 we present explanatory rules for propositional connectives. We assume these rules not to distinguish between formulas that are equivalent by associativity and commutativity of conjunction and disjunction, substitution of variables, and change of orders of identical quantifiers,[13] and to only apply to *c-sequents*.[14] Also their application is conditioned by the following restrictions.

Definition 2.11. We assume the application of explanatory propositional rules[15] to obey the following restrictions. For any formula of the form $C[F \circ G]$:

- if $\circ = \wedge$ and $SC_{F \circ G}(C) \neq \varnothing$, then the rule $\circ_1$ can be applied if, only if, $C[F \circ G]$ has PUS or NES; the rule $\circ_2$ can be applied if, and only if, $C[F \circ G]$ has NES.

- if $\circ = \wedge$ and $SC_{F \circ G}(C) = \varnothing$, then the rule $\circ_1$ can always be applied; the rule $\circ_2$ can be applied if, and only if, $C[.]$ has a negative polarity.

- if $\circ = \vee$ and $SC_{F \circ G}(C) \neq \varnothing$, then the rule $\circ_1$ can be applied if, only if, $C[F \circ G]$ has NUS or PES; the rule $\circ_2$ can be applied if, and only if, $C[F \circ G]$ has PES.

[13] See, Genco (2024).

[14] As it will become clear in the sequel, the choice of restricting to c-sequents renders problematic their extension to first-order rules. Although this condition can be relaxed, we prefer to adopt it in the present paper to limit the intricacy of the formalization.

[15] Reading the rules bottom-up.

- if $\circ = \vee$ and $SC_{F \circ G}(C) = \varnothing$, then the rule $\circ_1$ can always be applied; the rule $\circ_2$ can be applied if, and only if, $C[.]$ has a positive polarity.

For any formula of the form $C[\neg(F \circ G)]$:

- if $\circ = \wedge$ and $SC_{\neg(F \circ G)}(C) \neq \varnothing$, then the rule $\neg\circ_1$ can be applied if, only if, $C[\neg(F \circ G)]$ has NUS or PES; the rule $\neg\circ_2$ can be applied if, and only if, $C[\neg(F \circ G)]$ has PES.

- if $\circ = \wedge$ and $SC_{\neg(F \circ G)}(C) = \varnothing$, then the rule $\neg\circ_1$ can always be applied; the rule $\neg\circ_2$ can be applied if, and only if, $C[.]$ has a positive polarity.

- if $\circ = \vee$ and $SC_{\neg(F \circ G)}(C) \neq \varnothing$, then the rule $\neg\circ_1$ can only be applied if, only if, $C[\neg(F \circ G)]$ has PUS and NES; the rule $\neg\circ_2$ can be applied if, and only if, $C[\neg(F \circ G)]$ has NES.

- if $\circ = \vee$ and $SC_{\neg(F \circ G)}(C) = \varnothing$, then the rule $\neg\circ_1$ can always be applied; the rule $\neg\circ_2$ can be applied if, and only if, $C[.]$ has a negative polarity.

We now comment on these explanatory rules, which extend those presented in Poggiolesi (2018); Genco (2021). Indeed, whilst in Poggiolesi (2018); Genco (2021) explanatory rules were formulated for a propositional language, here they involve quantifiers. Most importantly, whilst the explanatory rules proposed in Poggiolesi (2018); Genco (2021) could only operate on the main connective of the formula under consideration, just like standard logical rules, one of the main innovations of this paper is that explanatory rules, differently from logical rules, can operate inside first-order contexts. This novelty provides us with explanatory rules that are much more powerful than those so far introduced in the literature on explanation or grounding (e.g. see Fine (2012); Millson and Strasser (2019)). This strength aligns with the strength standardly associated with explanations.

Each of the explanatory rules is supposed to capture cases where the premisses are the total and immediate reasons for the conclusions. In Section 5, we prove that this is indeed the case. Here our remarks are at the more intuitive level. Note first that 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 single rule for 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 formulas from one side of the sequent to another. In other words, one cannot explain the truth of $\neg F$, from the falsity of F. 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, because of the aspects mentioned above (when introducing contexts) and which are discussed at more length in Poggiolesi (2016b), the connective of negation must be carefully treated in an explanatory context; this is why the converse of a formula (see Definition 2.3) is used in the rules $\neg\circ_1$ and $\neg\circ_2$.

Let us now turn to those rules that do not involve conditions: i.e., rules $\neg\neg, \circ_1$ 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. Note however that application of rules deep inside formulas involves

some limitations to preserve an adequate notion of explanation. Let us illustrate this on some paradigmatic examples. The following is an instance of rule $\neg\circ_1$:

$$\frac{\Rightarrow \neg p \quad \Rightarrow \neg q}{\Rightarrow \neg(p \vee q)}\ \neg\circ 1$$

The rule can be applied since $\circ$ is a disjunction and the context is empty. Thanks to the rule $\neg\circ_1$, we can explain, totally and immediately, the formula $\neg(p \vee q)$ by means of the formulas $\neg p$ and $\neg q$, which are its reasons. The rule matches example 1.1. in Figure 1, and thus stands as an adequate instance of the rule. Let us also analyze an application of the same rule on the formula $\neg(\neg p \vee \neg q)$. In this case we have:

$$\frac{\Rightarrow p \quad \Rightarrow q}{\Rightarrow \neg(\neg p \vee \neg q)}\ \neg\circ 1$$

The rule can be applied since $\circ$ is a disjunction and the context is empty. Thanks to the rule $\neg\circ_1$, we can explain, totally and immediately, the formula $\neg(\neg p \vee \neg q)$ by means of the formulas p and q, which are its reasons. In particular, note that the rule provides as reasons $(\neg p)^*$ and $(\neg q)^*$, and that, by Definition 2.3 (definition of converse), $(\neg p)^*$ corresponds to p and $(\neg q)^*$ corresponds to q. The rule thus faithfully reflects the intuition discussed above Definition 2.3.

Let us now move to the following instance of the rule $\circ_1$:

$$\frac{\Rightarrow \forall x(Zx \rightarrow Nx) \quad \Rightarrow \forall x(SNx \rightarrow Nx)}{\Rightarrow \forall x((Zx \vee SNx) \rightarrow Nx)}\ \circ 1$$

The rule can be applied since $\circ$ is a disjunction, the scope of the context relative to the formula $Zx \vee SNx$ is not empty, and the formula in a context at issue has NUS (see Definition 2.11). Thanks to the rule $\circ_1$, we can explain, totally and immediately, the formula $\forall x((Zx \vee SNx) \rightarrow Nx)$ by the formulas $\forall x(Zx \rightarrow Nx)$ and $\forall x(SNx \rightarrow Nx)$, which represent the reasons why it is true. The rule matches example 1.4. of Figure 2 and thus stands as a an adequate instance of the rule.

Finally, consider the following formula $\forall x(Nx \rightarrow Ex \vee Ox)$, which can be seen as formalizing the sentence "every natural number is either odd or even." Suppose that one focuses on the disjunction and would like to apply a rule on it. Since the disjunction occurs with a positive polarity inside an universal quantifier that bounds variables both in E and O, none of the rules of the calculus can be applied to it. But this again matches our intuitions, as it would be incorrect to claim that because every natural number is even and every natural number is odd, then every natural number is either even or odd.[16]

Let us now move to the rules which involve 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 (2016b)). 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

[16]Note that this does not involve that there is no reason at all which explains the truth of this formula. Such reason(s) might be found by operating on the universal quantifier, or by relying on the concepts involved. The only point we aim to emphasize here is that whatever the reasons of the formula $\forall x(Nx \rightarrow Ex \vee Ox)$ are, they cannot be found be operating on the disjunction of the formula.

Figure 5: Explanatory first-order rules.

$$\frac{M \Rightarrow N, C[Fy]}{M \Rightarrow N, C[\odot x.Fx]}\ \odot_1 \qquad \frac{M \Rightarrow N, C[\odot x.Fx], C[Ft]}{M \Rightarrow N, C[\odot x.Fx]}\ \odot_2$$

$$\frac{M \Rightarrow N, C[F^*y]}{M \Rightarrow N, C[\neg(\odot x.Fx)]}\ \neg\odot_1 \qquad \frac{M \Rightarrow N, C[\neg(\odot x.Fx)], C[F^*t]}{M \Rightarrow N, C[\neg(\odot x.Fx)]}\ \neg\odot_2$$

where in $\odot_1$ and $\neg\odot_1$ y does not occur free in M nor in N.

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 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, in the scope of no quantifier (see Definition 2.11). 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 explanatory step matches what we have just been discussing and thus stands as an adequate instance of the rule.

Finally, we make two important remarks about all explanatory rules. The first concerns the fact that these rules do not distinguish between formulas that are equivalent by associativity and commutativity of conjunction and disjunction, substitution of variables, and change of orders of identical quantifiers. Consider for example the formulas $\forall x\forall y((Px \vee Ry) \rightarrow (Px \wedge Ry))$ and $\forall y\forall x((Px \vee Ry) \rightarrow (Ry \wedge Px))$, which are equivalent by change of orders of identical quantifiers and commutativity of conjunction. We can reasonably consider that these two formulas are explained by the same (multiset of) total and immediate reasons, for example $\forall x\forall y(Px \rightarrow (Px \wedge Ry)), \forall x\forall y(Ry \rightarrow (Px \wedge Ry))$, but also $\forall y\forall x(Px \rightarrow (Px \wedge Ry)), \forall y\forall x(Ry \rightarrow (Px \wedge Ry))$, and also $\forall y\forall x(Px \rightarrow (Ry \wedge Px)), \forall y\forall x(Ry \rightarrow (Ry \wedge Px))$, and so on. The explanatory rule reflects this feature.[17]

The second remark concerns the contexts in which the explanatory rules operate. When the explanatory rule has only one premise, like the rule $\neg\neg$, then the rule can be applied in any context $C[.]$. On the contrary, when the rule has two premisses, as all the other explanatory rules, attention needs to be paid to the quantifiers that bound variables occurring

[17]Note that this is a relevant property of the system which has strong connections with the so-called ground-theoretic, or factual, equivalence, so far only analyzed at the propositional level (e.g. see Correia (2016)), but also with the vast literature regarding the nature of proofs, such as proof-nets, or combinatorial proofs, (e.g. see Girard et al. (1989)).

in the formula on which the rule operates. Generally speaking, let $C[F]$ be the formula in the context on which the rule operates, and let $F = G \circ G'$. Then if $SC_{G \circ G'}(C)$ is not empty (namely if several quantifiers, under the scope of which $G \circ G'$ lies, bound the same variables in G as they do in G'), then the quantifiers need to be uniform,[18] and there needs to be a correspondence between quantifiers and main connective of the formula that lies in the scope of the quantifiers. If there is no variable common to G and G' that is bound by a quantifier, under the scope of which F lies, then quantifiers can be mixed. Technically, these restrictions depend on very well-known laws concerning the distributivity of quantifiers over connectives;[19] philosophically, as we have seen with the examples above, these restrictions guarantee that our rules formalize adequately the notion of (conceptual) explanations.

In Figure 5, we propose explanatory rules for quantifiers. Since we are providing a logical theory of conceptual explanations in first-order logic, we believe that we get a more elegant and harmonious overall theory if explanatory rules for quantifiers are displayed. However, it is worth emphasizing that in all the examples of conceptual explanations in the literature (e.g. see examples 1.2.-1.6.) quantifiers are typically left untouched, the explanation occurring inside them (hence motivating the use of contexts).

The explanatory rules for quantifiers are finitary rules, which we take to be a proof-theoretical desirable feature. Explanatory rules for quantifiers, like the explanatory rules for propositional connectives, extend inferential intuitions concerning the universal and the existential quantifiers at the explanatory level. Roughly speaking, the rule for the universal quantifier explains this quantifier by using the *eigenvariable*,[20] i.e. it explains why any object x has a property A via the fact that if one picks a random object y, y has the property A. This seems to correspond to what happens in mathematical contexts, where if a mathematician aims to explain why all triangles have a certain property, she will not work with all triangles, rather she will pick a triangle with no particular assumption on it - this is what the eigenvariable stands for - and prove that that triangle enjoys the property at issue. Since no particular assumption was invoked, she can generalize the explanation to all triangles.[21] The rule for the existential quantifier explains this quantifier via one of its instances; however, in order for the premisses of this rule to be the reasons of its conclusion, the existential itself needs to be repeated in the premisses. This move is analogous to that adopted in the rules $\forall L$ and $\exists R$ of the classical sequent calculus for first-order logic, e.g. see Troelstra and Schwichtenberg (1996). In the classical sequent calculus, the formula is repeated in the premise of the rule to make the rule invertible. In the case of the explanatory rule for the existential, the repetition of the formula serves to avoid the occurrence of conditions, which would be infinite. In other words, if one wants to keep the rule finitary, in the case of existential quantifier, conditions need to be given up: they are substituted by the repetition of the formula. This is a simple and useful technical device, although it may not be very satisfactory from a conceptual point of view. Another deep, yet more complicated, solution for the case of the existential is provided in Genco et al. (2021).

Finally, we also assume explanatory rules for quantifiers not to distinguish between formulas which are equivalent by associativity and commutativity of conjunction and disjunction, substitution of variables, and change of orders of identical quantifiers. Moreover, their application is conditioned on the following restrictions.

Definition 2.12. There is no restriction on the application of the rules $\odot_1$ and $\neg\odot_1$. We

[18] Namely they either need to be all universal quantifiers in contexts with positive polarity and existential in contexts with negative polarity, or vice-versa.

[19] E.g., see Casari (1997).

[20] See Troelstra and Schwichtenberg (1996).

[21] Note that analogous ideas have already been introduced and discussed in Genco et al. (2021).

assume the application of explanatory first-order rules $\odot_2$ and $\neg\odot_2$[22] to obey the following restrictions:

- rule $\odot_2$ can be applied on a formula of the form $C[\odot x.Fx]$ if, and only if, $C[.] \in \mathcal{P}$ and $\odot = \exists$, or $C \in \mathcal{N}$ and $\odot = \forall$.

- rule $\neg\odot_2$ can be applied on a formula of the form $C[\neg\odot x.Fx]$ if, and only if, $C[.] \in \mathcal{P}$ and $\odot = \forall$, or $C \in \mathcal{N}$ and $\odot = \exists$.

We will call $\mathbf{Gcl}^E$ the sequent calculus composed by the rules of Figures 3 and 4, whilst we will call $\mathbf{Gcl}^{EQ}$ the sequent calculus composed by the rules of Figures 3, 4 and 5. Here, since we focus on closed formulas, we will mainly deal with the calculus $\mathbf{Gcl}^E$, leaving results concerning $\mathbf{Gcl}^{EQ}$ for future research.

Definition 2.13. A (standard) *derivation* in $\mathbf{Gcl}^E$ is a finite (upward-growing) tree with a single root. The nodes of the tree are labelled by sequents and the sequents at the top nodes which are not initial sequents form the multiset S (that may be empty). For each non-terminal node, its label is connected with the labels of the immediate predecessor nodes by one of the rules of Figure 3 (except rule *cw*). The root of the tree is the conclusion of the whole derivation and in case its label is the sequent $M \Rightarrow N$, we say that there exists a derivation of $M \Rightarrow N$ from the assumptions S, in symbols $S \vdash_{\mathbf{Gcl}^E} M \Rightarrow N$. In case S is empty, we say that $M \Rightarrow N$ is a theorem, in symbols $\vdash_{\mathbf{Gcl}^E} M \Rightarrow N$.

Let $S, S', ...$ be multisets of c-sequents. Then, $(S)^* = \{(M \Rightarrow N)^* : M \Rightarrow N \in S\}$, where the converse of a c-sequent, $(M \Rightarrow N)^*$, corresponds to the formulas $\bigwedge M, \bigwedge N^*$.

Definition 2.14. A *total and mediate formal explanation* in $\mathbf{Gcl}^E$ is a finite (upward-growing) tree with a single root. The nodes of the tree are labelled by c-sequents or c-sequents with a bar; the c-sequents at the top nodes on the right side of the bar form the multiset S, whilst the c-sequents at the top nodes on the left side of the bar form the multiset S' (which could be empty). For each non-terminal node, its label is connected with the labels of the immediate predecessor nodes by one of the rules of Figure 4. The root of the tree is the conclusion of the explanation and is totally explained by the c-sequents S under the conditions $(S')^*$: in symbols $S' \mid S \Vdash_m M \Rightarrow N$. A *total and immediate formal explanation* from S to $M \Rightarrow N$, under conditions $(S')^*$ (which might be empty), in symbols $S' \mid S \Vdash M \Rightarrow N$, is any total and mediate formal explanation with one rule.

In the calculus $\mathbf{Gcl}^E$ it is thus possible to construct standard derivations, that formalize the notion of proof, but also formal explanations, which formalize the notion of explanatory proof, or conceptual explanation. Finally, in the calculus $\mathbf{Gcl}^E$ it is also possible to construct mixed derivations, which are standard derivations that might contain explanatory steps.[23]

Definition 2.15. A *mixed derivation* in $\mathbf{Gcl}^E$ is a derivation where also explanatory rules, as well as the rule *cw*, might have been applied. We use the notation $S' \mid S \vdash^{\star}_{\mathbf{Gcl}^E} M \Rightarrow N$, where both S and S' might be empty, to denote a mixed derivation in the calculus $\mathbf{Gcl}^E$.

[22] Reading the rules bottom-up.

[23] Note that these three types of logical objects - namely standard derivations, formal explanations, and mixed derivations - have been first introduced in Genco (2021), and then further analysed in Genco (2024).

3 Eliminability of the explanatory rules in the calculus $\mathbf{Gcl}^E$

In the previous section, we have introduced the calculus $\mathbf{Gcl}^E$ which is a calculus composed by the sequent calculus $\mathbf{Gcl}^+$ plus explanatory rules for the classical propositional connectives. In $\mathbf{Gcl}^E$ not only one can construct standard derivations (denoted by the symbol $\vdash$), but also derivations with explanatory steps (denoted by the symbol $\vdash^\star$), as well as formal explanations (denoted by the symbol $\Vdash_m$). As concerns standard derivations, $\mathbf{Gcl}^E$ is equivalent to **Gcl** and it keeps the same properties as **Gcl**.[24]

Lemma 3.1. *For any sequent $M \Rightarrow N$, $\vdash_{\mathbf{Gcl}} M \Rightarrow N$ if, and only if, $\vdash_{\mathbf{Gcl}^E} M \Rightarrow N$.*

Proof. Straightforward. □

Lemma 3.2. *The structural rules of weakening and contraction are height-preserving admissible in $\mathbf{Gcl}^E$. The logical rules of $\mathbf{Gcl}^E$ are height-preserving invertible (and given a logical rule $\mathcal{R}$, we will call $\overline{\mathcal{R}}$ its inverse).*

Proof. The proof is the same as that developed in **Gcl**, see (Troelstra and Schwichtenberg, 1996, Ch. 3.5). □

As concerns explanations, we need to show that nothing can be explained that cannot be derived, i.e. explanatory rules serve to build *derivations with an explanatory power*, not to prove new theorems.[25] To this end, we show that any explanatory rule can also be replaced by several applications of the standard inferential rules.

Lemma 3.3. *For any multiset of sequents S and S', any sequent $M \Rightarrow N$, and for any mixed derivation d of $M \Rightarrow N$ from S and S', namely $S' \mid S \vdash^\star M \Rightarrow N$ which contains only one application of an explanatory rule, one can construct a derivation d' with the same end-sequent from the same multiset of assumptions, namely $S \vdash M \Rightarrow N$.*

Proof. We reason by induction on the height of the derivation. We divide the explanatory rules into two groups: explanatory rules without conditions, namely $\neg\neg, \circ_1, \neg\circ_1$ and explanatory rules with conditions, namely $\circ_2, \neg\circ_2$. We start analyzing the rules of the first group. Suppose that the main formula of the premise of the explanatory rule is of the form $C[F]$. We apply on the context $C[.]$ as many rules $\overline{R}$ as necessary to unfold the context itself and reach the formula F.[26] Once arrived to F, given that explanatory rules do not distinguish between formulas which are FOL-equiv, we might need to further apply $\overline{R}$-rules to further decompose the formulas composing $C[.]$. We then apply the standard logic rules to get from F, or any formula FOL-equivalent to F, to the desired conclusion, and then we also use the logical rules to reconstruct the context $C[.]$. Here is a simple example of the procedure. Consider the following instance of the explanatory rule $\neg\neg$:

$$\dfrac{M \Rightarrow N, \forall x(Qx \wedge Px \rightarrow Rx)}{M \Rightarrow N, \forall x(Px \wedge Qx \rightarrow \neg\neg Rx)}\ \neg\neg$$

We obtain the desired result in the following way:

[24] Note that this also straightforwardly holds for $\mathbf{Gcl}^{EQ}$.

[25] This is an important feature which has been underlined in many papers, e.g. see Betti (2010); Poggiolesi (2016b).

[26] If C is empty, this first step of the procedure can be skipped.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{M \Rightarrow N, \forall x(Qx \wedge Px \rightarrow Rx)}{M \Rightarrow N, Qc \wedge Pc \rightarrow Rc}\,\forall R}{Qc \wedge Pc, M \Rightarrow N, Rc}\,\rightarrow R}{Pc, Qc, M \Rightarrow N, Rc}\,\wedge L}{\neg Rc, Pc, Qc, M \Rightarrow N}\,\neg L}{Pc, Qc, M \Rightarrow N, \neg\neg Rc}\,\neg R}{Pc \wedge Qc, M \Rightarrow N, \neg\neg Rc}\,\wedge L}{M \Rightarrow N, Pc \wedge Qc \rightarrow \neg\neg Rc}\,\rightarrow R}{M \Rightarrow N, \forall x(Px \wedge Qx \rightarrow \neg\neg Rx)}\,\forall R$$

As for the rules of the second group, namely those explanatory rules with conditions, one needs to consider the mixed derivation d, which either starts from leafs containing conditions, or leafs not containing any condition, and might involve applications of the rule cw. We substitute the derivation d with a derivation d' with no application of the rule cw, and in case d started from leafs containing conditions, we substitute them with the very same leafs but where all conditions have been erased. Then we continue the procedure as above. □

Lemma 3.4. *For any multiset of sequents S and S', any sequent $M \Rightarrow N$, and for any mixed derivation d of the form $S' \mid S \vdash^{\star} M \Rightarrow N$, one can construct a derivation d' from S to $M \Rightarrow N$, namely $S \vdash M \Rightarrow N$.*

Proof. By several applications of Lemma 3.3. □

Corollary 3.5. *For any multiset of sequents S and S', any sequent $M \Rightarrow N$, and for any formal explanation f of the form $S' \mid S \Vdash_m M \Rightarrow N$, one can construct a derivation d from S to $M \Rightarrow N$, namely $S \vdash M \Rightarrow N$.*

Proof. From Lemma 3.3. □

4 From reasons to conclusions

The main aim of this paper is to develop a logical theory of conceptual explanations. As noted, this involves addressing two central questions. The first - what kind of structure underlies conceptual explanations? - has been answered in Section 2 with the introduction of explanatory rules in the sequent calculus defining (formal) explanations. We use this section to answer the second question: what kind of features need to be satisfied for some formulas to count as the (total and immediate) reasons of another? As we have done for the first question, in order to answer this question, we will extend Poggiolesi's work; in particular, whilst Poggiolesi (2016b) proposes sufficient and necessary conditions to identify the grounds for a truth by only considering toy examples, here we will generalize her results to also encompass more refined cases (see Figures 1 and 2). In particular, working in first-order logic, we will put out those features that are sufficient and necessary to establish that, under certain conditions N, M are the total and immediate reasons of why A is true. The discussion will proceed into two main stages. First, we will introduce the features informally, and then move to the formal definitions.

The first feature that we need to consider in order to model the relation which links (total) reasons, or grounds, to their conclusion (i.e. the grounding relation) amounts to the widespread observation (e.g. see Jansson (2017); Kim (1994), Woodward (2004)), that this is a *dependence* relation. This dependence can, in its turn, be conveyed in the following

terms. In a relation that goes from the (total) reasons to their conclusion not only does the conclusion follow from its reasons, but it is also the case that if the reasons were modified somehow (under certain conditions), then this change would affect the conclusion. Translated into logical terms this becomes: not only it is the case that the conclusion logically follows from the (total) reasons, but also the negation of the conclusion needs to logically follow from the negation of some (even all) the (total) reasons (under certain conditions).

Let us consider this idea of dependency expressed in logical terms on the background of the example 1.3. (Figure 1) from the Introduction. It logically follows from their coding competences and teamwork skills that Stanford Computer Science graduates are ideal candidates for a particular tech job. However there seems to be more than just a logical consequence relation between these relata: indeed, if one of the premisses (or even both) were modified, this change would affect the conclusion. Suppose for example that Stanford Computer Science graduates do not have teamwork skills, then it follows that they no longer are ideal candidates for the tech job.

The conclusion is thus dependent on its reasons or grounds; however, this is not all. Indeed, any explanatory relation is asymmetric: there is a direction from what explains to what is explained. The dependency does not provide such a directionality. To see this clearly, one can consider any case with a single reason. Example 1.5. above (see Figure 2) perfectly fits this type of scenario: we indeed have that a property of circles is explained by an unique reason, namely a property of points. Note that these two properties can be proved to be equivalent, in line with our formulation of dependency. Yet, despite their equivalence and following Bolzano's intuitions, it is the property of points that explains the property of circles and not vice versa. We need to find an ingredient that determines this directionality or asymmetry. Poggiolesi[27] relies on an old and illustrious philosophical tradition[28] in identifying the missing ingredient as *complexity*: the simpler reasons explain the more complex conclusion, not the other way. Moreover, the increase in complexity in a grounding relation should be of a particular type:[29] the formulas by means of which a sentence is explained should correspond to a decomposition of the sentence itself. Although this insight is clear, deep as well as supported in the philosophical tradition, problems arise when we try to formalize it. The first notions that would seem to naturally serve the purpose are the standard notions of logical complexity and subformula; however, they turn out to be inadequate for explanations. Indeed, they face two main kinds of counterexample.[30] The first, concerning negation, can be illustrated by returning to example 1.1. (Figure 1). As already discussed, the (total and immediate) reasons of the formula $\neg(p \vee q)$ are $\neg p$ and $\neg q$. However $\neg p$ and $\neg q$ taken together are neither less complex[31] nor subformulas of $\neg(p \vee q)$, according to the standard notions. The second type of counterexample arises when considering cases like 1.2-1.6, namely cases where, as already underlined, one needs to look deep inside a formula. Consider in particular example 1.4. (Figure 2.).[32] In this case the (total and immediate) reasons of the formula $\forall x((Zx \vee SNx) \rightarrow Nx)$ are the formulas $\forall x(Zx \rightarrow Nx)$ and $\forall x(SNx \rightarrow Nx)$. Again, these two latter formulas are not logically less

[27] On this point, see further work in Poggiolesi (2018, 2024).

[28] E.g. see Betti (2010); Detlefsen (1988).

[29] See Rumberg (2013).

[30] Other features of logical complexity and subformulas that are not adequate for an explanatory framework will naturally emerge during the discussion.

[31] Here we mean that the sum of the logical complexity of $\neg p$ and $\neg q$ is lower than the logical complexity of $\neg(p \vee q)$.

[32] As already emphasized, although examples 1.2., 1.3. as well as 1.5. and 1.6., all display an analogous logical form, their explanatory power rely on the use of concepts the analysis of which goes beyond the purpose of the present paper. Such an analysis can be however found in Poggiolesi (2024).

complex nor subformulas of the conclusion, according to the standard definitions of logical complexity and subformula. In particular, whilst the standard notion of subformula only allows to break a formula along its main connective, the present example vividly shows that the possibility of breaking a formula from the inside, i.e. by breaking a connective that is not the main one and leaving the rest of the formula untouched, should also be taken into account.

We take these examples not as challenges to the idea that explanation involves an increase in complexity, but rather as a demonstration that standard notions of logical complexity and subformula are not fit for this purpose. We will thus enrich the notions of complexity and subformula to adapt them for an explanatory framework. In particular, we will first introduce the notion of *e-complexity*, that extends the standard notion of logical complexity by providing a more explanatory compelling way of counting connectives. Consequently, and in accordance with the new notion of e-complexity, we will define another relation of subformula, called *e-subformula*, that extends the standard notion of subformula in a way which is adequate for an explanatory framework.

Definition 4.1. Let $A \in \mathcal{WF}$, the e-complexity of A, $ecm(A)$, is defined in the following way:

- $ecm(Pt) = ecm(\neg Pt) = 0$
- $ecm(\neg\neg A) = ecm(A) + 1$
- $ecm(A \circ B) = ecm(\neg(A \circ B)) = ecm(A) + ecm(B) + 1$
- $ecm(\odot xAx) = ecm(\neg \odot xAx) = ecm(Ax) + 1$

Definition 4.1 relies on a previous definition of complexity for an explanatory framework provided in Poggiolesi (2016b) and extends it to the first-order level. Let us briefly illustrate the main insight behind it. It is a notion that basically aims at depicting a hierarchy of first-order formulas that lies in the background of the explanatory framework. Since in an explanatory framework, one goes from truth to truths, e-complexity tracks relationships among the truths expressed by the formulas, if they were true. In the cases of conjunction, disjunction and quantifiers, e-complexity coincides with the standard notion of logical complexity. If, for example, A and B express truths, then the truth expressed by $A \wedge B$ is obtained from the previous truths using a single operation. Thus conjunction increases by the sum of the e-complexity of A and that of B. Analogously, if A expresses a truth, then the truth expressed by $\forall xAx$ is obtained from the previous truth using a single operation. Thus the universal quantifier increases by one the e-complexity of the formula it is applied to. Things are more subtle for the case of negation. Let us see this first with the case of literals. Since (at most) one of Pc and $\neg Pc$ will express a truth, then only one of these formulas will ever be an object of an explanatory hierarchy. Thus, there seems to be no reason to count $\neg Pc$ as more complex than Pc: $\neg Pc$ can no longer be seen as constructed from Pc, since if one is true, the other is false. We should rather consider them as two formulas on the same level and this is precisely what e-complexity does. Analogous reasoning can be applied to the e-complexity of more complex formulas like $A \wedge B$ and $\neg(A \wedge B)$, or $\forall xAx$ and $\neg(\forall xAx)$. We can no longer count the complexity of $\neg(A \wedge B)$ as the complexity of $A \wedge B$ plus one, as standard logical complexity does, since if $\neg(A \wedge B)$ is true, then $A \wedge B$ is false and thus it cannot be constructed from it. We should rather think of $A \wedge B$ and $\neg(A \wedge B)$ as two formulas that lie at the same level of an explanatory hierarchy and thus have the

same e-complexity. The exception is the case of double negation, where the negation counts since $ecm(\neg\neg A) = ecm(A) + 1$. But this is in harmony with what has been said up to now: $\neg\neg A$ and A may both express truths, and thus the former can be seen as constructed from the latter by means of a single operation.

Note also that thanks to the notion of e-complexity, we can look at the relation between a formula A, and its converse A^* (see Definition 2.3), under a novel light. Indeed each formula and its converse are such that their conjunction corresponds to a contradiction and they have the same e-complexity.[33] Note also that in an explanatory framework one may work with contexts (see Definitions 2.4) and formulas in contexts (see Definitions 2.5). The e-complexity of contexts, and formulas in contexts can be defined as follows.

Definition 4.2. We define the e-complexity of a context $ecm(C[.]) = ecm(C[Pc])$ for any predicate P and constant c in $\mathcal{L}$.

Definition 4.3. We define the e-complexity of a formula in context, $ecm(C[F])$ as a pair of numbers (m, n) such that $m = ecm(C[.])$ and $n = ecm(F)$. Accordingly, given the formulas in a context $C_1[F_1], ..., C_k[F_k]$ and $D[G]$, if $ecm(C_1[F_1]) = (m, n_1)$, ..., $ecm(C_k[F_k]) = (m, n_k)$ and $ecm(D[G]) = (m, n)$, where $n = n_1 + ... + n_k + 1$, then $C_1[F_1], ..., C_k[F_k]$ will be said to be *immediately less g-complex* than $D[G]$.

We now move to our new notion of subformula, that will be called *e-subformula*, and that will work in parallel with the notion of e-complexity (just as logical complexity and subformula do). There are three main ideas that motivate the new notion of e-subformula. The first idea is related to the aforementioned fact that in an explanatory framework relations amongst formulas might involve connections that go deep inside formulas themselves. The standard subformula only connects formulas by looking at the main connective; we will enrich it by also allowing to look at connectives inside the formulas. As a consequence, we will use again the notions of context, and formula in a context. The second and third ideas are linked to the novel way of counting the complexity of a formula. Consider formulas F and E which are equivalent by associativity and commutativity of conjunction and disjunction, change of orders of identical quantifiers, and substitution of variables. Not only are F and E logically equivalent, they also are equivalent from an explanatory point of view. Indeed, E and F convey the same "state of affairs," and occupy the same place in the explanatory hierarchy, i.e. they have the same e-complexity. Hence if F is a subformula of F', then E should be as well. We will render this feature by closing the relation of e-subformula under associativity and commutativity of conjunction and disjunction, change of orders of identical quantifiers, and substitution of variables.

Note that this sort of reasoning also applies to any formula F and its converse F^*. Although F and F^* are of course not equivalent, yet they share a deep relation: they convey the same"state of affairs" and they occupy the same place in the explanatory hierarchy, i.e. they have the same e-complexity. Either F is true or F^* is, yet they represent the two sides of the same coin. As a result, whenever a formula F' is a e-subformula of a formula F, its converse will be too.

Now that we have clarified the main insights behind the new notion of *e-subformula*, we introduce it formally via the following definitions.

Definition 4.4. Given the formulas F and G of $\mathcal{L}$, we say that F is *FOL-equiv* to G if, and only if, F can be obtained from G by associativity and commutativity of conjunction and disjunction, substitution of variables, and change of orders of identical quantifiers.

[33] Whilst a formula and its negation are such that their conjunction corresponds to a contradiction, but they do not necessarily have the same e-complexity.

Definition 4.5. Given a context $C[.]$ of $\mathcal{L}$, we say that $C[.]$ is *FOL-equiv* to $D[.]$ if, and only if, for any predicate P and any constant $c \in \mathcal{L}$, $C[Pc]$ is FOL-equiv to $D[Pc]$.

Definition 4.6. For any pair of formulas F and G of $\mathcal{L}$, we say that $F \cong G$ if, and only if, F is FOL-equiv to G or F is FOL-equiv to G^*.

Definition 4.7. For any pair of contexts $C[.]$ and $D[.]$ of $\mathcal{L}$, we say that $C[.] \cong D[.]$ if, and only if, for any predicate P and any constant c in $\mathcal{L}$, $C[Pc]$ is FOL-equiv to $D[Pc]$ or $C[Pc]$ is FOL-equiv to $(D[Pc])^*$.

Definition 4.8. For any pair of multisets M and N of formulas of $\mathcal{L}$, such that $M = \{C_1[F_1], ..., C_n[F_n]\}$ and $N = \{D_1[G_1], ..., D_n[G_n]\}$, we say that $M \cong N$, if, and only if, $F_1 \cong G_1$, ..., $F_n \cong G_n$ and $C_1 \cong D_1$, ..., $C_n \cong D_n$.

Definition 4.9. For any pair of formulas in contexts $C[F]$ and $D[G]$ of $\mathcal{L}$, we say that $D[G]$ is a e-subformula of $C[F]$ if, and only if, $C[.] \cong D[.]$, and:

- $F \cong G$,
- $F \cong \neg\neg F'$ and G is a e-subformula of F',
- $F \cong F' \circ F''$ and G is a e-subformula of F' or G is a e-subformula of F'',
- $F \cong \odot x F'$ and G is a e-subformula of $F'(t/x)$ for all t free for x in F'.

The notion of *immediate e-subformula* is analogous to that of immediate subformula.

Definition 4.10. M is a multiset of *distinguished immediate e-subformulas* of $C[F]$, if, and only if:

- $M \cong \{C[F']\}$ and $F \cong \neg\neg F'$,
- $M \cong \{C[F'], C[F'']\}$ and $F \cong (F' \circ F'')$,
- $M \cong \{C[\odot x F']\}$ and $F \cong F'(t/x)$, for all t free for x in F'.

Note that the distinguished immediate e-subformulas of $C[F]$ are always immediately less e-complex than $C[F]$ according to Definition 4.3, so that the notion of e-complexity and e-subformula go hand in hand.

We now have all the ingredients to formally define the necessary and sufficient conditions which establish when, under certain conditions N, formulas M count as the total and immediate reasons of a formula A.

Definition 4.11. For any finite multisets of $\mathcal{CF}$ $M = \{D_1[G_1], ..., D_m[G_m]\}$ and $N = \{C_1[F_1], ..., C_n[F_n]\}$(which could be empty), and for any $\mathcal{CF}$ $C[F]$, M is a *total and immediate formal reason* of $C[F]$ under the condition that N^*, in symbols $N \mid M \Vdash C[F]$, if, and only if, for any $E[.]$ such that $SC(E) = SC(C)$ and $E[.] \in \mathcal{P}$ if, and only if, $C[.] \in \mathcal{P}$, we have:

1. $E[G_1], ..., E[G_m] \models E[F]$,
2. for some non empty (possibly non proper) submultiset M' of M, such that $M' = \{D_{k1}[G_{k1}], ..., D_{kr}[G_{kr}]\}$, we have that $(E[F_1])^*, ..., (E[F_n])^*, (E[G_{k1}])^*, ..., (E[G_{kr}])^*, M^-/E \models (E[F])^*$.

3. $N \cup M$ is a multiset of distinguished immediate e-subformulas of $C[F]$.

where $M^- = M - M'$ and $M^-/E = \{E[G_z] : D_z[G_z] \in M^-\}$.

Definition 4.11 represents the formal counterpart of the features discussed in this section. Conditions 1. and 2. are meant to capture the dependency of the relation. Obviously this dependence holds amongst the formulas (in contexts) at issue, independently from the contexts these formulas belong to. For this reason, it is relevant to demand the dependency for any context $E[.]$ whose scope and polarity are the same as that of the formula to be explained.[34] Condition 3. amounts to the directionality or asymmetry of the explanatory relation at issue: this is conveyed via the new notion of e-subformula.[35]

Let us evaluate some relations of reasons to related conclusions which emerge from this definition. Consider the formula $\neg(p \vee q)$ that we have discussed in the example 1.1. above (see Figure 1), for which, as noted $\neg p, \neg q$ are the total and immediate reasons. Definition 4.11 matches this intuition. Indeed $\neg(p \vee q)$ is a classical logical consequence of $\neg p$ and $\neg q$, but it is also the case that if we modify a subset of the reasons and we consider, say, p and $\neg q$, instead of $\neg p$, $\neg q$, it logically follows that $p \vee q$. Finally, $\{\neg p, \neg q\}$ is the multiset of distinguished immediate e-subformulas of $\neg(p \vee q)$.

Let us now turn to the formula $\forall x((Zx \vee SNx) \rightarrow Nx)$ from the example 1.4 (see Figure 2), whose total and immediate reasons are the formulas $\forall x(Zx \rightarrow Nx)$ and $\forall x(SNx \rightarrow Nx)$. Definition 4.11 again agrees with this intuition. For any context $E[.]$, such that $SC(E) = \forall x$ and $E \in \mathcal{N}$, we have that $E[Zx \vee SNx]$ logically follows from $E[Zx]$ and $E[SNx]$; but it is also the case that if we modify the reasons, so we consider, say $E[Zx]$ and $(E[SNx])^*$, then it logically follows that $(E[Zx \vee SNx])^*$. Finally, $\{\forall x(Zx \rightarrow Nx), \forall x(SNx \rightarrow Nx)\}$ is a multiset of distinguished immediate e-subformulas of $\forall x((Zx \vee SNx) \rightarrow Nx)$.

Finally, consider the formula $\forall x(Nx \rightarrow Ex \vee Ox)$ mentioned in Section 2, that could be seen as formalizing the sentence "for any x, if x is a natural number, then it is an odd or an even number." Although the formulas $\forall x(Nx \rightarrow Ex)$ - for any x if x is a natural number, then it is even - and $\forall x(Nx \rightarrow Ox)$ - for any x if x is a natural number, then it is odd - are both e-subformulas of $\forall x(Nx \rightarrow Ex \vee Ox)$, it would be rather strange to think of them as its reasons, if only because they are false. Definition 4.11 confirms this intuition: it can be easily checked that condition 2. does not hold between the well-formed closed formula $\forall x(Nx \rightarrow Ex \vee Ox)$ and the formulas $\forall x(Nx \rightarrow Ex)$ and $\forall x(Nx \rightarrow Ox)$: in particular, from $(\forall x(Nx \rightarrow Ex))^*$ and $(\forall x(Nx \rightarrow Ox))^*$ (or even the converse of just one of them), it does not follow that $(\forall x(Nx \rightarrow Ex)$ and $\forall x(Nx \rightarrow Ox))^*$.

We can extend the definition of total and immediate formal reasons-conclusion (or equivalently, the definition of the total and immediate formal grounding relation) to total and mediate formal reasons-conclusion in the following way.

Definition 4.12. For any multisets of $\mathcal{CF}$ M and N (which could be empty), and for any $\mathcal{CF}$ F, under the condition that N^*, M is a *total and mediate formal reason* of F, $N \mid M \Vdash_m F$, if, and only if:

[34] Although it arose in a different framework, a similar dependence relation has been investigated in Humberstone (2013).

[35] Note that Definition 4.11 picks up the main insights of Definition 6.1 given in Poggiolesi (2016b) and extends it in three different ways. On the one hand, and as already underlined, dependency between the reason and their conclusion is extended to rely on any context (with the same polarity and scope of the formula to be explained) and g-complexity (see Poggiolesi (2016b)) is replaced by e-complexity. These two modifications lead to a notion of total and immediate formal reasons of a generality incomparably wider than that which can be found in Poggiolesi (2016b). On the other hand, (i) the second condition of Definition 6.1. of Poggiolesi (2016b) has been slightly modified, and (ii) the possibility of having multiple conditions and not just one has been added. The necessity of both (i) and (ii) were motivated and discussed in Poggiolesi and Francez (2021).

- $N \mid M \Vdash F$, or
- $N' \mid M' \Vdash G$ and $N'' \mid G, M'' \Vdash_m F$, where $M' \cup M'' = M$, and $N' \cup N'' = N$.

5 Explanatory rules provide all, and only, relations from (total) reasons to conclusion

In this section we prove that the answers offered to our two motivating questions – concerning the structure of explanations and the relationship between reasons and conclusions – are consistent. More specifically we show that a conclusion follows from reasons (in the sense of $\Vdash$) if, and only if, there is an explanation from the former to the latter (i.e. $\Vdash$ holds). In particular, Theorem 5.7 establishes that if a rule is explanatory then its premisses represent the total and immediate reasons of it conclusion, according to Definition 4.11. Theorem 5.14 will prove that if some formulas count as the total and immediate reasons of a certain conclusion (according to Definition 4.11), then there exists an explanatory rule which convey this relation. Other lemmas and definitions serve to prove these main theorems.

Lemma 5.1. *The following rules are admissible in the calculus* ***Gcl****:*

when $C[.] \in \mathcal{P}$*:* *when* $C[.] \in \mathcal{N}$*:*

$$\frac{C[F_i], M \Rightarrow N}{C[F_1 \wedge F_2], M \Rightarrow N}\ {}^{\wedge 1} \qquad \frac{M \Rightarrow N, C[F_i]}{M \Rightarrow N, C[F_1 \vee F_2]}\ {}^{\vee 1} \qquad \frac{M \Rightarrow N, C[F_i]}{M \Rightarrow N, C[F_1 \wedge F_2]}\ {}^{\wedge 2} \qquad \frac{C[F_i], M \Rightarrow N}{C[F_1 \vee F_2], M \Rightarrow N}\ {}^{\vee 2}$$

$$\frac{C[F_i^*], M \Rightarrow N}{C[\neg(F_1 \vee F_2)], M \Rightarrow N}\ {}^{\neg\vee 1} \qquad \frac{M \Rightarrow N, C[F_i^*]}{M \Rightarrow N, C[\neg(F_1 \wedge F_2)]}\ {}^{\neg\wedge 1} \qquad \frac{M \Rightarrow N, C[F_i^*]}{M \Rightarrow N, C[\neg(F_1 \vee F_2)]}\ {}^{\neg\wedge 2} \qquad \frac{C[F_i^*], M \Rightarrow N}{C[\neg(F_1 \wedge F_2)], M \Rightarrow N}\ {}^{\neg\vee 2}$$

where $i = \{1, 2\}$.

Proof. We prove in detail the admissibility of the rules $\wedge 1$ and $\wedge 2$ by induction on the construction of the context $C[.]$, and subinduction on the height of the derivation of the premise of the rule. The admissibility of the other rules can be proved analogously.

We distinguish cases according to the form of $C[.]$. If $C[.] = [.]$, then from the premise $F_i, M \Rightarrow N$ we obtain the desired result thanks to the rule $\wedge L$.

If $C[.] \neq [.]$, then we distinguish cases according to the last applied rule $\mathcal{R}$ on $C[F_i], M \Rightarrow N$ and on $M \Rightarrow N, C[F_i]$. (i) A rule $\mathcal{R}$ has been applied on either M or N. In this case we apply the inductive hypothesis on the height of the derivation, and then by re-applying $\mathcal{R}$ we get the desired result. (ii) A rule $\mathcal{R}$ has been applied on $C[F_i]$ in the sequent $C[F_i], M \Rightarrow N$ (the case where $\mathcal{R}$ has been applied on $C[F_i]$ in the sequent $M \Rightarrow N, C[F_i]$ is analogous). We distinguish the following sub-cases according to the form of $C[.]$.

$$\frac{E, D[F_i], M \Rightarrow N}{E \wedge D[F_i], M \Rightarrow N} \rightsquigarrow^{36} \frac{E, D[F_1 \wedge F_2], M \Rightarrow N}{E \wedge D[F_1 \wedge F_2], M \Rightarrow N}$$

$$\frac{E, M \Rightarrow N \quad D[F_i], M \Rightarrow N}{E \vee D[F_i], M \Rightarrow N} \rightsquigarrow \frac{E, M \Rightarrow N \quad D[F_1 \wedge F_2], M \Rightarrow N}{E \vee D[F_1 \wedge F_2], M \Rightarrow N}$$

[36] The symbol $\rightsquigarrow$ means: the premise of the right side is obtained by inductive hypothesis on height of the derivation of the premise of the left side.

When $C[F_i] = \forall x D[F_i]$, we proceed as follows:

$$\dfrac{\forall x D[F_i], D[F_i], M \Rightarrow N}{\forall x D[F_i], M \Rightarrow N} \quad \rightsquigarrow \quad \dfrac{\dfrac{\forall x D[F_1 \wedge F_2], D[F_i], M \Rightarrow N}{\forall x D[F_1 \wedge F_2], D[F_1 \wedge F_2], M \Rightarrow N}\ {\scriptstyle i.h.}}{\forall x D[F_1 \wedge F_2], M \Rightarrow N}$$

where in the derivation on the right, in the top inference, *i.h.* stands for the fact that the inductive hypothesis on the construction of the context $C[.]$ allows us to infer the sequent $\forall x D[F_1 \wedge F_2], D[F_1 \wedge F_2], M \Rightarrow N$.

Suppose finally that $C[F_i]$ is of the form $\neg D[F_i]$[37] and that the sequent $\neg D[F_i], M \Rightarrow N$ has been obtained from the sequent $M \Rightarrow N, D[F_i]$ by means of the rule $\neg L$. Then we consider the sequent $M \Rightarrow N, D[F_i]$ and we apply (since now $D[.] \in \mathcal{N}$) the rule $\wedge_2$ obtaining the desired result.

□

Lemma 5.2. *For any pair of formulas $F, \neg\neg F \in \mathcal{CF}$, it holds that:*

$$C[F] \Vdash C[\neg\neg F]$$

Proof. By induction on the construction of $C[.]$. If $C[.]$=[.], then it is trivial. If $C[.] \neq [.]$, then we need to distinguish cases. However, since F and $\neg\neg F$ are logically equivalent, it is straightforward to check that it holds for any case. □

Definition 5.3. Given $G, G', F \in \mathcal{CF}$:

we write $G, G' \div F$ when $G, G' \models F$ and $G^*, G'^* \models F^*$.

we write $G \mid G' \div F$ when $G' \models F$ and $G^*, G'^* \models F^*$.

we write $< G > G' \div F$ when $G'^* \models F^*$ and $G, G' \models F$.

Lemma 5.4. *For any $G, G', F \in \mathcal{CF}$:*

$$G \mid G' \div F \text{ if, and only if, } < G^* > G'^* \div F^*$$

Proof. Straightforward. □

Lemma 5.5. *For any context $C[.]$ that has positive polarity and for any formula $G, G', F \in \mathcal{CF}$ such that $F \in \{G \wedge G', \neg(G \vee G')\}$, then it holds that:*

(a) if $G, G' \div F$, then $C[G], C[G'] \div C[F]$,

(b) if $< G > G' \div F$, then $< C[G] > C[G'] \div C[F]$,

where if $SC_F(C)$ is not empty, then $C[F]$ has PUS (see Definition 2.10).

For any context $C[.]$ that has negative polarity and for any formula $G, G', F \in \mathcal{CF}$ such that $F \in \{G \wedge G', \neg(G \vee G')\}$, then it holds that:

(c) if $G, G' \div F$, then $C[G], C[G'] \div C[F]$,

[37] The case where $C[F_i]$ is of the form $\neg\neg D[F_i]$ is clearly analogous.

(d) if $< G > G' \doteqdot F$*, then* $C[G] \mid C[G'] \doteqdot C[F]$*,*

where if $SC_F(C)$ *is not empty, then* $C[F]$ *has NES (see Definition 2.10).*

Proof. We prove (a)-(d) by (a common) induction on the the construction of $C[.]$. **We start from (a)**. If $C[.] = [.]$, then it is trivial. Suppose $C[.] \neq [.]$, then we distinguish cases according to the form of C. We have (i) $C = \neg\neg D[.]$, (ii) $C = E \wedge D[.]$,[38] (iii) $C = E \vee D[.]$,[39] (iv) $C = \forall x D[.]$, (v) $C = \exists x D[.]$, (vi) $C = \neg D[.]$.

(i). It is straightforward.

(ii). Suppose $G, G' \doteqdot F$. By i.h., one obtains $D[G], D[G'] \doteqdot D[F]$. In order to get the desired result, we exploit the sequent calculus **Gcl** in the following way:[40]

$$\dfrac{\dfrac{\dfrac{D[G], D[G'] \Rightarrow D[F] \quad E, E \Rightarrow E}{E, D[G], E, D[G'] \Rightarrow E \wedge D[F]}\ {\wedge R'}}{E, D[G], E \wedge D[G'] \Rightarrow E \wedge D[F]}\ {\wedge L}}{E \wedge D[G], E \wedge D[G'] \Rightarrow E \wedge D[F]}\ {\wedge L} \qquad \dfrac{\dfrac{\dfrac{D[F] \Rightarrow D[G], D[G'] \quad E \Rightarrow E}{E, D[F] \Rightarrow D[G], E \wedge D[G']}\ {\wedge R'} \quad E \Rightarrow E}{E, E, D[F] \Rightarrow E \wedge D[G], E \wedge D[G']}\ {\wedge R'}}{E, D[F] \Rightarrow E \wedge D[G], E \wedge D[G']}\ {CL}}{E \wedge D[F] \Rightarrow E \wedge D[G], E \wedge D[G']}\ {\wedge L}$$

From $E \wedge D[G], E \wedge D[G'] \vdash E \wedge D[F]$ by completeness of **Gcl**, one gets $E \wedge D[G], E \wedge D[G'] \models E \wedge D[F]$. From $E \wedge D[F] \vdash E \wedge D[G] \vee E \wedge D[G']$ by completeness of **Gcl**, and the symbol of converse (see Definition 2.3), one gets $(E \wedge D[G])^*, (E \wedge D[G'])^* \models (E \wedge D[F])^*$. Thus we have $E \wedge D[G], E \wedge D[G'] \doteqdot E \wedge D[F]$.

(iii). Analogously to (ii).

(iv) In this case we further distinguish sub-cases according to whether (iva) $SC_F(C) \neq \varnothing$, or (ivb) $SC_F(C) = \varnothing$. We start by analyzing (iva). We further distinguish this case, according to the form of F. We thus have (iva$'$) $F = Gx \wedge G'x$, and (iva$''$) $F = \neg(Gx \vee G'x)$.

(iva$'$). By i.h., one obtains $D[Gx], D[G'x] \doteqdot D[Gx \wedge G'x]$. One gets the desired result, exploiting rule $\wedge 1$ of Lemma 5.1, as well as the sequent calculus **Gcl**, in the following way:[41]

$$\dfrac{\dfrac{\dfrac{D[Gx], D[G'x] \Rightarrow D[Gx \wedge G'x]}{\forall x D[Gx], D[G'x] \Rightarrow D[Gx \wedge G'x]}\ {\forall L'}}{\forall x D[Gx], \forall x D[G'x] \Rightarrow D[Gx \wedge G'x]}\ {\forall L'}}{\forall x D[Gx], \forall x D[G'x] \Rightarrow \forall x D[Gx \wedge G'x]}\ {\forall R} \qquad \dfrac{\dfrac{\forall x D[Gx] \Rightarrow \forall x D[Gx]}{\forall x D[Gx \wedge G'x] \Rightarrow \forall x D[Gx]}\ {\wedge 1}}{\forall x D[Gx \wedge G'x] \Rightarrow \forall x D[Gx], \forall x D[G'x]}\ {WR}$$

From $\forall x D[Gx], \forall x D[G'x] \vdash \forall x D[Gx \wedge G'x]$ by completness of **Gcl** one gets $\forall x D[Gx], \forall x D[G'x] \models \forall x D[Gx \wedge G'x]$. From $\forall x D[Gx \wedge G'x] \vdash \forall x D[Gx] \vee \forall x D[G'x]$ by completeness of **Gcl**, and the symbol of converse (see Definition 2.3), one gets $(\forall x D[Gx])^*, (\forall x D[G'x])^* \models (\forall x D[Gx \wedge G'x])^*$. Thus we have $\forall x D[Gx], \forall x D[G'x] \doteqdot \forall x D[Gx \wedge G'x]$.

[38] The case $C = D[.] \wedge E$ is analogous.

[39] The case $C = D[.] \vee E$ is analogous.

[40] For the sake of simplicity, we use the multiplicative version of the rule $\wedge R$, as well as the rule of contraction on the left side of the sequent, which are both admissible rules in the calculus **Gcl**.

[41] For the sake of simplicity, we use the version of the rule $\forall L$ and $\exists R$ without the repetition of the quantifier, as well as the weakening on the right. These rules are admissible in the calculus **Gcl**.

(iva″). Analogously to (iva′) by using the rule $\neg \vee 1$, whose admissibility has been shown in Lemma 5.1.

(ivb) We should further distinguish cases depending on whether (ivb′) the quantifier bounds no variable at all or a variable in G but not in G' (or vice versa); or (ivb″) the quantifier bounds some variable in $D[.]$. In the former case, the procedure is straightforward, in the latter case one proceeds as in (iva).

(v) In this case, since we are dealing with an existential, we have that $SC_F(C) = \varnothing$. Given that, there are mainly two sub-cases to treat. (va) The quantifier bounds a variable in either G or G', (vb) the quantifiers bound some variable in $D[.]$.

(va) By i.h., one obtains $D[Gx], D[G'] \doteqdot D[F]$ (we assume the variable x to occur in G and not in G′, the inverse situation can be treated analogously). In order to get the desired result, we exploit the sequent calculus **Gcl** in the following way:

$$\dfrac{\dfrac{\dfrac{D[Gx], D[G'] \Rightarrow D[Fx]}{D[Gx], D[G'] \Rightarrow \exists x D[Fx]}\,\exists R'}{\exists x D[Gx], D[G'] \Rightarrow \exists x D[Fx]}\,\exists L}{\exists x D[Gx], \exists x D[G'] \Rightarrow \exists x D[Fx]}\,\exists L \qquad \dfrac{\dfrac{\dfrac{D[Fx] \Rightarrow D[Gx], D[G']}{D[Fx] \Rightarrow \exists x D[Gx], D[G']}\,\exists R}{D[Fx] \Rightarrow \exists x D[Gx], \exists x D[G']}\,\exists L}{\exists x D[Fx] \Rightarrow \exists x D[Gx], \exists x D[G']}\,\exists L$$

From $\exists x D[Gx], \exists x D[G'] \vdash \exists x D[Fx]$ by completness of **Gcl** one gets $\exists x D[Gx], \exists x D[G'] \models \exists x D[Fx]$. From $\exists x D[Fx] \vdash \exists x D[Gx] \vee \exists x D[G']$ by completeness of **Gcl**, and the symbol of converse (see Definition 2.3), one gets $(\exists x D[Gx])^*, (\exists x D[G'])^* \models (\exists x D[Fx])^*$. Thus we have $\exists x D[Gx], \exists x D[Gx] \doteqdot \exists x D[Fx]$.

(vb) By i.h., one obtains $D[G], D[G'] \doteqdot D[F]$. In order to get $(\exists x D[G])^*, (\exists x D[G'])^* \models (\exists x D[F])^*$, we proceed as in the case above, namely exploiting the sequent calculus, the completeness of the sequent calculus and the symbol of converse. In order to get the other side, namely $\exists x D[G], \exists x D[G'] \models \exists x D[F]$, we start from $D[G], D[G'] \vdash D[F]$ and we apply as many rules $\overline{R}$ as necessary to divide the context D (relative to the formulas $D[G]$ and $D[G']$), into two parts: the part which contains the variable which will be bound by the existential and the part without it. We thus either get (i) $D', D''[G], D', D''[G'] \vdash D[F]$, if the connective linking D' and D'' in D is a conjunction; (ii) $D' \vdash D[F]$ and $D''[G], D''[G'] \vdash D[F]$, if the connective linking D' and D'' in D is a disjunction. In the former case we proceed as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{D', D''[G], D', D''[G'] \Rightarrow D[F]}{D', D''[G], D''[G'] \Rightarrow D[F]}\,CL}{D[G], D''[G'] \Rightarrow D[F]}\,\wedge L{+}\mathcal{R}}{D[G], D''[G'] \Rightarrow \exists x D[F]}\,\exists R'}{\exists x D[G], D''[G'] \Rightarrow \exists x D[F]}\,\exists L}{\exists x D[G], D', D''[G'] \Rightarrow \exists x D[F]}\,WL}{\exists x D[G], D[G'] \Rightarrow \exists x D[F]}\,\wedge L{+}\mathcal{R}}{\exists x D[G], \exists x D[G'] \Rightarrow \exists x D[F]}\,\exists L$$

In the latter case we proceed as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{D' \Rightarrow D[F]}{D', D''[G'] \Rightarrow D[F]}\, WL \quad D''[G], D''[G'] \Rightarrow D[F]}{D[G], D''[G'] \Rightarrow D[F]}\, \vee L{+}\mathcal{R}}{D[G], D''[G'] \Rightarrow \exists x D[F]}\, \exists R' \qquad \dfrac{\dfrac{D' \Rightarrow D[F]}{D' \Rightarrow \exists x D[F]}\, \exists R'}{\exists x D[G], D' \Rightarrow \exists x D[F]}\, WL}{\exists x D[G], D''[G'] \Rightarrow \exists x D[F]}\, \exists L \qquad \vee L{+}\mathcal{R}$$

$$\dfrac{\exists x D[G], D[G'] \Rightarrow \exists x D[F]}{\exists x D[G], \exists x D[G'] \Rightarrow \exists x D[F]}\, \exists L$$

(vi) Assuming $G, G' \doteqdot F$, we apply (c) getting $D[G], D[G'] \doteqdot D[F]$, where F has a negative polarity. However, by logic, this is equivalent to $\neg D[G], \neg D[G'] \doteqdot \neg D[F]$, which is the desired result and where $D[.]$ has a positive polarity.

The cases (b)-(d) can be treated analogously to case (a).

□

Lemma 5.6. *For any context $C[.]$ that has positive polarity and for any formula $G, G', F \in \mathcal{CF}$ such that such that $F \in \{G \vee G', \neg(G \wedge G')\}$, then it holds that:*

(a) if $G, G' \doteqdot F$, then $C[G], C[G'] \doteqdot C[F]$,

(b) if $G \mid G' \doteqdot F$, then $C[G] \mid C[G'] \doteqdot C[F]$,

where if $SC(C)_F$ is not empty, then $C[F]$ has PES (see Definition 2.10).

For any context $C[.]$ that has negative polarity and for any formula $G, G', F \in \mathcal{CF}$, such that $F \in \{G \vee G', \neg(G \wedge G')\}$, then it holds that:

(c) if $G, G' \doteqdot F$, then $C[G], C[G'] \doteqdot C[F]$,

(d) if $G \mid G' \doteqdot F$, then $< C[G] > C[G'] \doteqdot C[F]$,

where if $SC(C)_F$ is not empty, then $C[F]$ has NUS (see Definition 2.10).

Proof. The proof is analogous to the proof of Lemma 5.5. □

Theorem 5.7. *(Soundness) For any multisets of sequents S', S (where S' is possibly empty), and sequent $M \Rightarrow N$,*

$$if\ S' \mid S \Vdash M \Rightarrow N,\ then\ (S')^{\tau} \mid (S)^{\tau} \Vdash \bigwedge M \rightarrow \bigvee N$$

where $(S')^{\tau}$, $(S)^{\tau}$ are the standard translation of the multisets of sequents into multisets of formulas.

Proof. In order to prove the theorem, we should check the validity of each explanatory rule of Figure 4. The validity of the rule $\neg\neg$ follows from Lemma 5.2. We prove the validity of rule $\circ_1$. The validity of the other rules can be proved analogously.

Consider the rule $\circ_1$ applied on a formula of the form $C[F \wedge G]$, where $C[.]$ has a positive polarity. Clearly, it holds that $F, G \doteqdot F \wedge G$. But, then by Lemma 5.5, we have $\bigwedge M \rightarrow$

$\bigvee N \vee C[F], \bigwedge M \rightarrow \bigvee N \vee C[G] \doteqdot \bigwedge M \rightarrow \bigvee N \vee C[F \wedge G]$, where if $SC_{F \wedge G}(\bigwedge M \rightarrow \bigvee N \vee C)$ is not empty, then $\bigwedge M \rightarrow \bigvee N \vee C[F \wedge G]$ has PUS. Actually for Lemma 5.5 again, we have that, for any context $E[.]$ it holds that $E[F], E[G] \doteqdot E[F \wedge G]$, where if $SC_{F \wedge G}(E)$ is not empty, then $E[F \wedge G]$ has PUS. Finally, $\{\bigwedge M \rightarrow \bigvee N \vee C[F], \bigwedge M \rightarrow \bigvee N \vee C[G]\}$ is a multiset of immediate distinguished e-subformulas of $\bigwedge M \rightarrow \bigvee N \vee C[F \wedge G]$ (also thinking of FOL-equivalent formulas). Hence we have the desired result.

Consider the rule $\circ_1$ applied on a formula of the form $C[F \wedge G]$ where $C[.]$ has a negative polarity. Then the reasoning is the same as above and it thus crucially relies on Lemma 5.5.

Consider the rule $\circ_1$ applied on a formula of the form $C[F \vee G]$ where $C[.]$ has a positive polarity. Then the reasoning is the same as above, except that one needs to use Lemma 5.6.

Consider the rule $\circ_1$ applied on a formula of the form $C[F \vee G]$ where $C[.]$ has a negative polarity. Then the reasoning is the same as above, except that one needs to use Lemma 5.6. □

Corollary 5.8. *For any multiset of sequents S', S (where S' is possibly empty), and sequent $M \Rightarrow N$,*

$$if\ S' \mid S \Vdash_m M \Rightarrow N,\ then\ (S')^\tau \mid (S)^\tau \Vdash_m \bigwedge M \rightarrow \bigvee N$$

where $(S')^\tau$, $(S)^\tau$ are the standard translation of the multisets of sequents into multisets of formulas.

Proof. From Theorem 5.7. □

Definition 5.9. For any context $C[.]$, we define the related quantifiers-only-context $Qo(C)[.]$, in the following way:

- if $C[.] \in \mathcal{P}$, then $Qo(C)[.] = SC(C)[.]$
- if $C[.] \in \mathcal{N}$, then $Qo(C)[.] = SC(C)[.]^*$

where $[.]^*$ stands for $\neg[. \wedge \top]$.[42]

Lemma 5.10. *Let $Qo(C)[.]$ be the quantifiers-only-context related to $C[.]$, then:*

$$SC(Qo(C)) = SC(C)$$

Proof. Straightforward from Definition 5.9. □

Lemma 5.11. *For any multisets of $\mathcal{CF}$ M and N (which could be empty), and for any $\mathcal{CF}$ $C[F]$,*

$$if\ N \mid M \Vdash C[F],\ then\ Qo(N) \mid Qo(M) \Vdash Qo(C)[F]$$

where for any multiset of closed formulas P, $Qo(P) = \{Qo(E)[G] \mid E[G] \in P\}$.

[42]The need of writing $\neg[. \wedge \top]$, instead of the simpler $\neg[.]$, is motivated by the way contexts have been defined, i.e. it is not always possible to have a context of the form $\neg[.]$. On the other hand, in this specific case, we need to consider a context with a negative polarity.

Proof. By Definition 4.11. Indeed, since by such Definition, the relation should hold for any context $E[.]$ such that $SC(E) = SC(C)$ and $E[.] \in \mathcal{P}$ if, and only if, $C[.] \in \mathcal{P}$, then it will also hold for those contexts where quantifiers are at the top of the formulas under consideration, as these contexts satisfy both conditions above by the way they are constructed. As for the third clause of Definition 4.11, note that by the way subformulas are defined (see Definition 4.9) if $N \cup M$ is a multiset of distinguished immediate e-subformulas of $C[F]$, so is $Qo(N) \cup Qo(M)$ with respect to $Qo(C)[F]$. Indeed, $Qo(N)$, $Qo(M)$ and $Qo(C)[F]$ are all obtained in the same way from N, M and $C[F]$, respectively.

□

Definition 5.12. For any quantifier-only-context $Qo(C)[.]$, we say that $Qo(C)[.]$ is:

a *positive universal* if, and only if, $Qo(C)[.] = \forall x_1, ...\forall x_n[.]$, where $n \geqslant 0$.

a *negative universal* if, and only if, $Qo(C)[.] = \forall x_1, ...\forall x_n[.]^*$, where $n \geqslant 0$.

a *positive existential* if, and only if, $Qo(C)[.] = \exists x_1, ...\exists x_n[.]$, where $n \geqslant 0$.

a *negative existential* if, and only if, $Qo(C)[.] = \exists x_1, ...\exists x_n[.]^*$, where $n \geqslant 0$.

Lemma 5.13. *For any multisets of $\mathcal{CF}$ M and N (which could be empty), and for any $\mathcal{CF}$ $C[F]$,*

$$\textit{if } Qo(N) \mid Qo(M) \Vdash Qo(C)[F] \textit{ then } (N)^\delta \mid (M)^\delta \Vdash \Rightarrow C[F]$$

where for any multiset of $\mathcal{CF}$ M, $M^\delta = \{\Rightarrow E[C] \mid E[C] \in M\}$.

Proof. We proceed by distinguishing cases based on the form of $Qo(C)[.]$ and F.

[-] $Qo(C)[.]$ might be such that: (i) it is a positive universal; (ii) it is a negative universal; (iii) it is a positive existential; (iv) it is a negative existential; (v) $Qo(C)[.]= SC(C)[.]$, where $SC(C)$ corresponds to any finite sequence of universal and existential quantifiers that is not empty and is neither of the type $\forall x_1, ..., \forall x_n$, nor of the type $\exists x_1, ..., \exists x_n$; (vi) $Qo(C)[.]= SC(C)[.]^*$, where $SC(C)$ corresponds to any finite sequence of universal and existential quantifiers that is not empty and is neither of the type $\forall x_1, ..., \forall x_n$, nor of the type $\exists x_1, ..., \exists x_n$.

[-] F can be of the following form: (a) $\neg\neg G$; (b) $G \wedge G'$; (c) $G \vee G'$; (d) $\neg(G \wedge G')$; (d) $\neg(G \vee G')$; (e) $\forall x Gx$; (f) $\neg\forall x Gx$; (g) $\exists x Gx$; (h) $\neg\exists x Gx$.

We check in detail the combinations of (i)-(vi) with (a), (b) and (e). The other combinations can be treated analogously.

1. We combine (i)-(vi) with (a). In each case, we have that $Qo(C)[G] \Vdash Qo(C)[\neg\neg G]$; at the syntactic level the explanatory rule $\neg\neg$ gives us $\Rightarrow C[G] \Vdash \Rightarrow C[\neg\neg G]$, as required.

2. We combine (i) with (b). We have that $Qo(C)[G], Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$. At the syntactic level, thanks to the explanatory rule $\circ_1$, we obtain $\Rightarrow C[G], \Rightarrow C[G'] \Vdash \Rightarrow C[G \wedge G']$, as required.

3. We combine (iv) with (b). We have that $Qo(C)[G], Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$, $Qo(C)[G] \mid Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$ and $Qo(C)[G'] \mid Qo(C)[G] \Vdash Qo(C)[G \wedge G']$. At the syntactic level, thanks to the explanatory rules $\circ_1$, $\circ_2$, we get $\Rightarrow C[G], \Rightarrow C[G'] \Vdash \Rightarrow C[G \wedge G']$, $\Rightarrow C[G] \mid\Rightarrow C[G'] \Vdash \Rightarrow C[G \wedge G']$ and $\Rightarrow C[G'] \mid\Rightarrow C[G] \Vdash\Rightarrow C[G \wedge G']$, as required.

4. We combine (ii) with (b). We distinguish between two sub-cases: in the first sub-case $SC_{G \wedge G'}(Qo(C))$ is not empty (this involves that there exists some variable that occurs both in G and G' that the quantifiers bound), whilst in the second case $SC_{G \wedge G'}(Qo(C))$ is empty (which means that there is no variable that occurs both in G and G' that the quantifiers bound). In the former sub-case, it is straightforward to check that no relation of total and immediate formal reasons can be established. In the latter sub-case, we have $Qo(C)[G], Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$, $Qo(C)[G] \mid Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$ and $Qo(C)[G'] \mid Qo(C)[G] \Vdash Qo(C)[G \wedge G']$. At the syntactic level, the explanatory rules $\circ_1$, $\circ_2$ give us what required.

5. We combine (iii) with (b). We distinguish between two sub-cases: in the first sub-case $SC_{G \wedge G'}(Qo(C))$ is not empty, whilst in the second case $SC_{G \wedge G'}(Qo(C))$ is empty. In the former sub-case, it is straightforward to check that no relation of total and immediate formal reasons can be established. In the latter sub-case, we have $Qo(C)[G], Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$. At the syntactic level, the explanatory rule $\circ_1$ give us what required.

6. We combine (v) with (b). We distinguish between two sub-cases: in the first sub-case $SC_{G \wedge G'}(Qo(C))$ is not empty, whilst in the second sub-case $SC_{G \wedge G'}(Qo(C))$ is empty. In the latter sub-case, we have $Qo(C)[G], Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$. At the syntactic level, thanks to the explanatory rule $\circ_1$ we get what required. In the former sub-case, we need to further distinguish depending on the type of quantifiers $SC_{G \wedge G'}(Qo(C))$ contains. If $SC_{G \wedge G'}(Qo(C)) = \forall y_1, ...\forall y_n$, then, as before we have $Qo(C)[G], Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$ and the explanatory rule $\circ_1$ gives us what required. In all the other cases, no relation of total and immediate formal reason arises.

7. We combine (vi) with (b). We distinguish between two sub-cases: in the first sub-case $SC_{G \wedge G'}(Qo(C))$ is not empty, whilst in the second sub-case $SC_{G \wedge G'}(Qo(C))$ is empty. In the latter sub-case, we have $Qo(C)[G], Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$, $Qo(C)[G] \mid Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$ and $Qo(C)[G'] \mid Qo(C)[G] \Vdash Qo(C)[G \wedge G']$. At the syntactic level, thanks to the explanatory rules $\circ_1$, $\circ_2$, we get what desired. In the former sub-case, we need to further distinguish depending on the type of quantifiers $SC_{G \wedge G'}(Qo(C))$ contains. If $SC_{G \wedge G'}(Qo(C)) = \exists y_1, ...\exists y_n$, then, as before, we have $Qo(C)[G], Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$, $Qo(C)[G] \mid Qo(C)[G'] \Vdash Qo(C)[G \wedge G']$ and $Qo(C)[G'] \mid Qo(C)[G] \Vdash Qo(C)[G \wedge G']$. The explanatory rules $\circ_1$ and $\circ_1$ give us what required. In all the other cases, no relation of total and immediate formal reason arises.

8. We combine (i)-(vi) with (e), hence with a formula of the type $Qo(C)[\forall x Ax]$ It is easy to check that there is no *closed* e-subformula of $Qo(C)[\forall x Ax]$ such that it stands with $Qo(C)[\forall x Ax]$ in a relation of total and immediate reasons-conclusion. Hence, this case does not need to be further analyzed.

□

Theorem 5.14. *(Completeness) For any multisets of closed formulas N, N' (possibly empty), and formula $C[F]$,*

$$if\ N' \mid N \Vdash C[F],\ then\ (N')^{\delta} \mid (N)^{\delta} \Vdash \Rightarrow C[F]$$

Proof. From Lemmas 5.11 and 5.13. □

Corollary 5.15. *For any multisets of closed formulas N, N' (possibly empty), and formula $C[F]$,*

$$if\ N' \mid N \Vdash_m C[F],\ then\ (N')^\delta \mid (N)^\delta \Vdash_m \Rightarrow C[F]$$

Proof. From Theorem 5.14. □

6 Related work

Recent years have witnessed an increasing interest in the notion of explanation from a logical point of view. Here we mention some trends (the list is not exhaustive) that could be seen as related to the present work. A first trend is that which is taking place with some common machine learning classifiers, where recent research has aimed at identifying the reasons behind the classification of instances, and thus has proposed explanations for this type of decision, e.g. see Darwiche and Hirth (2023); Shih et al. (2018). By considering the examples taken into account in this wide and still flourishing literature, and in particular by dwelling on the crucial notion of *sufficient reason*, we are led to believe that there are strong analogies with our approach. If this analogy is well-founded, then it could prove fruitful in (at least) two ways. On the one hand, the semantics put forward by Darwiche and Hirth (2023) - which is mainly in terms of *prime implicants* - can be profitably employed to develop a semantic-approach for the present proposal that is for now mainly syntactic. On the other hand, the power of the explanatory sequent calculus introduced here could be implemented to enrich the computational part of the work developed in, e.g. Shih et al. (2018); it could also be used to extend their perspective, which to date concerns solely the propositional level, to a first-order language.

Another direction, related to Darwiche and Hirth (2023); Shih et al. (2018) is the work of Liu and Lorini (2023, 2022). In a nutshell, Liu and Lorini introduce the reasons behind the classification of instances in a (modal) language, and develop an axiomatic system as well as a semantics for the new connective. Since the approach we propose only lies at the metalinguistic level,[43] but is lacking for its linguistic counterpart, the work of Liu and Lorini (2023, 2022) can be seen as a useful source of inspiration.

A third trend that one might be tempted to consider lies within metaphysics, where there is a growing interest towards the notion of (metaphysical) grounding (e.g. see Fine (2012)). Although there exists several attempts to develop a logic of metaphysical grounding, these attempts are all based on toy examples (e.g. see Poggiolesi (2020a)). As a result, rules or axioms involved in these formalizations are much weaker and simpler than those introduced in this paper. The issue in this case will be that of verifying whether one can witness more intricate cases of metaphysical grounding, closer to the type that can be found in mathematics. If this is the case - as we suspect - then results of this paper can contribute to discussion in the metaphysical literature as well.

Last, but not least, another recent trend in the current literature concerns those explanations that are characterized by abductive reasoning, namely an inference to the best

[43] This is a further difference with Poggiolesi (2018), where the connective $\rhd$, for *because*, was introduced. Given the intricacy of the issue, we leave the analysis of the extension of the connective $\rhd$ for connecting formulas in contexts at the first-order level for future research.

explanation. In this framework there are (at least) two recent lines of work, one developed by Arieli et al. (2022); Millson and Strasser (2019), whilst the other by Piazza et al. (2023); Pulcini and Varzi (2021). Despite their difference, these works have a strong common feature, namely they both develop new sequent calculi were several different rules are proposed to account for abductive reasoning. We thus have three proof-theoretical formal frameworks dealing with different notions of explanations. As a consequence, the study of their relations could open up for a novel and interesting connection between conceptual and abductive explanations, both at the conceptual and at the logical level.

7 Conclusions

The word *explanation* is an umbrella term which covers several different notions, such as causal, non-causal or abductive explanations. In this paper we have focussed on conceptual explanations, namely some deductive explanations-why, which come from a long and illustrious tradition in philosophy, bear several analogies with causal explanations, but still deserve a thorough formal study. The main aim of this paper has been to take some first steps towards filling this gap, by the introduction of a logical theory of the notion of (conceptual) explanation and related relation from reasons to conclusion (i.e. grounding relation). We have accomplished this task by using and enriching the standard tools of proof theory, namely the sequent calculus for classical first-order logic. In particular we have added to the standard inferential rules explanatory rules, i.e., rules whose premisses represent the (total and immediate) reasons why their conclusion is true. By means of these rules we can construct formal explanations, which represent the formalization of the notion of (conceptual) explanation. Not only do we believe that this research provides a valuable contribution *per se*, in that it fills an important gap in the logical literature, but it also naturally opens up several directions for future research, such as the formalization of the notion of explanation in logics other than classical logic, the applications of conceptual explanations to related fields such as explainable AI, or to related notions of explanation. Finally, it also open up to the investigation of the value of explanatory rules in proof-theoretic semantics, e.g. see Francez (2015).

References

Arieli, O., Borg, A., Hesse, M., and Strasser, C. (2022). Explainable logic-based argumentation. *Computational Models of Argument*, 353:32–43.

Arieli, O. and Strasser, C. (2015). Sequent-based logical argumentation. *Argument and computation*, 6:73–99.

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

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

Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press, Cambridge.

Bolzano, B. (2014). *Theory of Science*. Oxford University Press, Oxford.

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

Casari, E. (1997). *Introduzione alla Logica*. UTET Press.

Correia, F. (2016). On the logic of factual equivalence. *Review of Symbolic Logic*, 9:103–122.

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

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

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

Francez, N. (2015). *Proof-theoretic semantics*. College Publication.

Gallow, J. D. (2021). The metaphysics of causation. In Zalta, N. E., editor, *The Stanford Encyclopedia of Philosophy*, pages 1–39. Stanford.

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

Genco, F. (2024). What stands between grounding rules and logical rules is the excluded middle. *Review of Symbolic Logic*, forthcoming:1–24.

Genco, F., Poggiolesi, F., and Rossi, L. (2021). Grounding, quantifiers and paradoxes. *Journal of Philosophical logic*, 36:1–34.

Girard, J.-Y., Lafont, Y., and Taylor, P. (1989). *Proofs and Types*. Cambridge University Press, Cambridge.

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

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

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

Humberstone, L. (2013). Replacement in logic. *Journal of Philosophical Logic*, 42:49–89.

Hunmean, P. (2010). Topological explanations and robustness in biological sciences. *Synthese*, 177:213– 245.

Jansson, L. (2017). Explanatory asymmetries, ground, and ontological dependence. *Erkenntnis*, 82:95–136.

Kim, J. (1994). Explanatory knowledge and metaphysical dependence. *Philosophical Issues*, 5:51–69.

Kment, B. (2021). Varieties of modality. In Zalta, N. E., editor, *The Stanford Encyclopedia of Philosophy*, pages 1–35. Stanford.

Kortabarria, M. and Giannotti, J. (2024). Scientific explanation as a guide to ground. *Synthese*, 203:1–32.

Lange, M. (2017). *Because Without Cause: Non-causal Explanations in Science and Mathematics*. Oxford University Press, Oxford.

Lewis, D. (1973). Causation. *Journal of Philosophy*, 70:556–567.

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

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

Mancosu, P., Poggiolesi, F., and Pincock, C. (2023). Mathematical explanation. In *Stanford Encyclopedia of Philosophy*, pages 1–43. Stanford.

Menzies, P. and Beebee, H. (2020). Counterfactual theories of causation. In Zalta, E., editor, *The Stanford Encyclopedia of Philosophy*, pages 1–49. Stanford.

Millson, J. and Strasser, C. (2019). A logic for best explanations. *Journal of Applied Non-Classical Logics*, 29:184–231.

Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge.

Piazza, M., Pulcini, G., and Sabatini, A. (2023). Abduction as deductive saturation: a proof-theoretic inquiry. *Journal of Philosophical Logic*, 52:1575–1602.

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

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

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

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

Poggiolesi, F. (2020a). Logics. In Raven, M., editor, *Routledge Handbook for Metaphysical Grounding*, pages 213–227. Routledge, New York.

Poggiolesi, F. (2020b). A proof-theoretical framework for several types of grounding. *Logique et Analyse*, 252:387–414.

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

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

Poggiolesi, F. and Francez, N. (2021). Toward a generalization of the logic of grounding. *Theoria*, 36:5–24.

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

Pulcini, G. and Varzi, A. (2021). Classical logic through refutation and rejection. In Fitting, M., editor, *Landscapes in Logic (Volume on Philosophical Logics)*, pages 1–45. College Publications.

Rumberg, A. (2013). Bolzano's concept of Abfolge against the background of normal proofs. *Review of Symbolic Logic*, 6:424–459.

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

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

Steiner, M. (1973). Mathematical explanations. *Philosophical Studies*, 34:135–151.

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

Woodward, J. (2004). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.