Explanations in logic

Francesca Poggiolesi

Francesca Poggiolesi
IHPST, UMR 8590
Paris, France
poggiolesi@gmail.com

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 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 the sequent calculus. 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.

1 Introduction

To explain phenomena in the world is a central human enterprise and one of the main goals 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 in the 20th century. In the light of this vast literature, it is useful to start with some orientation. First of all, the word *explanation* is an umbrella term to denote different types of activities, e.g. see Schurz (1999). If one kind of explanation is that which shows how to construct an Ikea furniture, another amounts to the clarification of the meaning of a symbol, and yet another type of explanation corresponds to the explanation of a new concept to a child. In this paper we will not take into account the world *explanation* in its generality, but just focus on the so-called *deductive explanations-why*, namely those explanations which have a deductive form and aim to clarify why a certain phenomenon occurs or why a certain proposition is true.

Amongst deductive explanations-why a central place is occupied by the so-called *causal explanations*, e.g. see Woodward (2004). Causal explanations are those explanations that track or can be identified with a causal relation. In other terms, causal explanations are explanations that explain why a certain phenomenon occurs by displaying the cause(s) that determine the phenomenon, which thus corresponds to the effect. Examples of causal explanations range from toy examples to bona fide scientific explanations. The argument which explains why there is a fire in the forest by evoking the cigarette lit in the forest (as well as the law of combustion) is an example of causal explanation, in that it explains by relying on the causal relation between the cigarette lit in the forest and the fire that it provoked. However, also the explanation of current climate damages which evokes our burning of fossil fuels (together with several physical and chemical laws linking burn of fossil fuels with climate damages) is another example of causal explanation in that it explains why climate change occurs by displaying one of its cause, namely the fact that we burn fossil fuels.

Only in the past decade or so philosophers have became increasingly aware of plenty of compelling examples of explanations-why that causal accounts cannot properly capture.

In physics as well as in mathematics or in metaphysics, several types of explanations arose that did not seem to rely on any causal mechanism: very naturally, the idea that causation although certainly being 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.[1]

Amongst the wide set of non-causal explanations, one might focus on *conceptual explanations*,[2] namely those explanations that track or can be identified with a (conceptual) grounding relation.[3] In other terms, conceptual explanations are those explanations which explain why a certain conclusion is true by displaying the reason(s) or ground(s)[4] why it is such, where the relation between such reasons and conclusion hold in virtue of the concepts that they contain. Examples of conceptual explanations range from toy examples to intricate ones. The argument which explains why a certain animal is a vixen by evoking that animal being a female as well as that animal 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 the animal is a vixen, and the relation between the reasons and their conclusion hold in virtue of the concepts - female, fox and vixen - that they contain. However, also the explanation of why Jane is the ideal candidate for the new professorship at a prestigious European university, which evokes the several qualities of Jane - she is a hard worker, she is talented, she has prestigious publications - (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 dwelling on the reason(s) why it is true.

Note that amongst conceptual explanations one should also count *mathematical explanations*,[5] namely those explanations that take the form of proofs in mathematics that not only show a theorem to be true, but also seem to provide the reason(s) why it is true. These mathematical explanations have been the object of several reflections by an illustrious tradition of scholars including Aristotle, Proclus, Leibniz, Arnauld and Nicole, Bolzano, Frege.[6] For the sake of clarity, we sketch an example of this type of mathematical explanation that comes from Bolzano (2014). Consider the theorem which states that 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[7] of this theorem that crucially relies on a property of points, namely the property which states that for any two points a and b, there always exists a third point c such that $l(ab) = l(bc) = l(ac)$. Following Bolzano, this proof is explanatory in that it relies on the grounding relation between the property of the points - the reason - and the property of the

[1] E.g. see Lange (2017); Reutlinger and Saatsi (2018).

[2] E.g. see Betti (2010); Schnieder (2006).

[3] On the links between conceptual explanations and grounding as they are adopted in this paper, see Poggiolesi and Genco (2023).

[4] In this paper, we use as synonymous the words "ground" and "reason." However, we do not take grounding to be a metaphysical relation as it is commonly assumed to be in the contemporary literature, e.g. see Fine (2012). In this paper we rather think of the notion of *conceptual ground*, which has been receiving an increasing attention recently, e.g., Betti (2010); Carrara and De Florio (2020); Smithson (2020).

[5] On the inclusion amongst conceptual explanations of mathematical explanations, see Betti (2010); Poggiolesi and Genco (2023).

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

[7] *Proof.* Consider the circle A with center a and radius ab. Since by definition a center is a point, then we have that there exists a point a. For the same reasoning applied to the circle B, we have that there exists a point b. But given a point a and a point b, there always exists a point c such that $l(ab) = l(bc) = l(ac)$ (where $l(xy)$ stands for the length of the segment xy). Hence we have a point c such that $l(ab) = l(bc) = l(ac)$. Since the distance between c and the centre of the circle A is the radius of A, and the same holds for B, c is a point where the two circles A and B intersect.

circles - the conclusion. In its turn, this grounding relation holds in virtue of the concepts the sentences it connects contain, namely the concepts of point, radius, circle. Hence this mathematical explanatory proof is a paradigmatic example of conceptual explanation in that it enjoys the several features of this type of explanation.

In sum, when it comes to deductive explanations why, there is a stringent parallel between the causal and the conceptual level: explanations belonging to these different frameworks share a similar structure, analogous features, several common properties. Although causal explanations are dominant in scientific inquiry and philosophy,[8] logic has been argued to have a problematic relationship with causality.[9] As a result, as far as we know, the study of (causal) explanation is a great absentee in the logic literature, a literature otherwise rich of formalizations with other central notions such as knowledge, belief, time, obligation and so on. Conceptual explanations, on the other hand, naturally invite a logical analysis, and this is precisely the aim of this paper, namely to *elaborate a logical theory of conceptual explanations.* Given the wide range of conceptual explanations, from toy examples to mathematical explanations, the theory will need to be rich enough to shed light on all these different cases of conceptual explanation. The elaboration of such a theory will have two consequences. On the one hand, it will allow us to introduce the notion of explanation in logic, where it has so far been absent; on the other hand, it will enlighten our understanding of conceptual explanations themselves.

Note that, in the contemporary literature, the line of research, which is the closest to that of this paper, is concerned with the logic of (metaphysical) grounding, e.g. see Correia and Schnieder (2012); Fine (2012). Indeed, because of the strict relation between conceptual explanations and grounding, there are several analogies between the two. However, three main novelties characterize this paper: (i) the focus on the notion of (conceptual) explanation rather than on the relation of grounding; (ii) the use of the sequent calculus for first-order logic, rather than the use of the natural deduction calculus; (iii) the attention dedicated to mathematical explanations that, being a quite elaborated type of conceptual explanation, naturally lead to the construction of a broader framework.

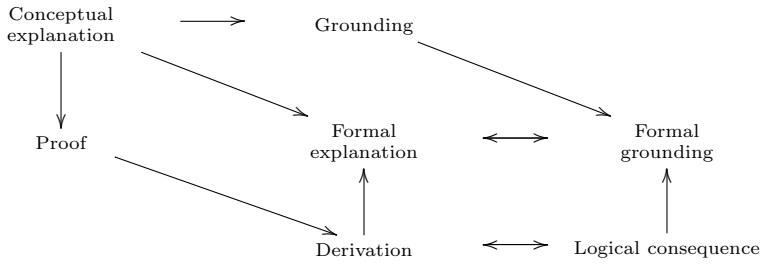
In order to reach our goal, we will organize the paper in the following way. In *Section 2* we will clarify the formal framework where we will develop our account along with some characteristics of conceptual explanations as well as the related grounding relation. Whilst in *Section 3* we will set out the conditions under which some formulas are the formal grounds of another, in *Section 4* we will formalize the notion of conceptual explanation, via the notion of formal explanation. *Section 5* will serve to prove soundness and completeness between formal grounding and formal explanation and in *Section 6* we will prove some results concerning our formal theory of explanation. In *Section 7* we will draw conclusions and sketch directions of future research.

2 Formal framework

In order to provide a formal account of conceptual explanation, let us start from an idea that is both ancient and central in the literature on (deductive) explanations, and which simply consists in seeing them as deductive *arguments* which, starting from true premisses - be them the causes or the grounds - explain a certain conclusion.[10] Of course not any deductive argument amounts to an explanation, but some of them do, namely those which

[8] E.g. see Hempel (1965); Salmon (1989).
[9] E.g. Scriven (1971). Only more recently has this trend been inverted, e.g. see Bareinboim et al. (2022).
[10] E.g. see Aristotle (1993); Hempel (1965, 1942).

Figure 1: General picture



have an explanatory power. The perspective that we will develop in these pages consists in a reformulation of this central idea along the following lines: explanations can be seen as *proofs* which, starting from true premisses, the grounds, not only prove that a certain conclusion is true, but also explain why it is such. This perspective not only naturally arises from the observation that proofs are deductive arguments, but is also supported by the fact that mathematical explanations, a notable subset of conceptual explanations, actually are proofs of mathematical theorems that show why 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 derivation. More precisely, we will introduce a metalinguistic relation that we will call *formal explanation*,[11] and which will represent the formal counterpart of conceptual explanation as well as a special case of the standard notion of derivation. The first task of this paper is to provide a definition of this new metalinguistic relation. However important this task might be, if we limited ourselves to it, we wouldn't be taking into account the general framework to which conceptual explanations belong. Indeed, when it comes to (at least a certain type of) explanations, these structures are strongly related to a relation: causal explanations with causality and conceptual explanations with grounding. Even derivations come equipped with a relation, namely the relation of logical consequence (to which, in several logics, they are shown to be equivalent to). Hence, to capture the framework of conceptual explanations in an adequate way, it is reasonable and relevant to also introduce the formal counterpart of the relation of grounding, that we will call *formal grounding*, and which we will conceive as a strengthening of the logical consequence relation (see Figure 1). The two metalinguistic relations of formal explanation and formal grounding will be the main objects of this paper, and we will denote them with the two symbols $\Vdash$ and $\Vdash\!\!\models$, respectively. In order to better understand what kind of features characterize these relations, let us introduce some further specifications.

First of all, we will analyze both symbols $\Vdash$ and $\Vdash\!\!\models$ in the language of classical first-order logic, which we introduce in the following way.

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

[11] Here we follow Poggiolesi (2018).

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 first-order classical logic.

Secondly, we introduce some notable distinctions that help identifying different types of explanations and associated relations. Here we focus on two, namely the distinction between *total/partial* explanations and grounding, and the distinction between *immediate/mediate* explanations and grounding.[12] A total explanation (grounding relation) is one which 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 (or ground) C is a total explanation (ground) of C. On the other hand, each proper sub-multiset of the total explanation (ground) of C is a partial explanation (ground) of C.

Let us now move to the distinction immediate/mediate. Whilst an immediate explanation (grounding relation) is one that involves a single explanatory step, i.e. that does not seem to be further reducible, a mediate explanation (grounding relation) includes several sequential immediate steps. In this paper we will first deal with the notions of total and immediate formal explanation, and total and immediate formal grounding,[13] and then generalize them both to the mediate case.

There exists a third distinction that is linked to the notion of total explanation (grounding) and that, once more, arises both in the causal and conceptual framework. To illustrate it, we start form the causal case, where it is most well-known and then move to the non-causal one. Consider the following notorious example.[14] Billy and Suzy throw rocks at a bottle. The glass 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 too and shattered the glass; Billy's rock sailed through the air. 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 someone's sister and Suzy, Jane's sister, has a girl. Jane's brother could have had a girl, but he does not. Hence Jane's brother having a girl is merely a potential reason of why Jane has a niece. Potential reasons are also central for total explanations: if Jane's brother had 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*.[15] So, for example, we will say that under the condition that Jane's brother does not have a girl, the total reason why Jane has a niece is that her sister has a child.

[12] Note that the distinctions total/partial and immediate/mediate not only hold for the notions of conceptual explanation and grounding, but also for causal explanation and causality, see Lewis (1973); Fine (2012); Schaffer (2016).

[13] Note that the notion of total and immediate formal grounding corresponds to what Poggiolesi (2016) calls *complete* and immediate formal grounding.

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

[15] Here we borrow vocabulary from Genco (2021).

3 Formal grounding relation

We start by considering the relation of formal grounding between a sentence A and a set of sentences M, under certain conditions M'. As we have already said, we see this relation as a special case of the classical logical consequence relation, namely A is a logical consequence of its reasons M. But what features distinguish a grounding relation from a logical consequence relation? To answer this question, we will rely on and extend some previous results which have been developed at the propositional level.[16] Also we will first illustrate the main ideas in an informal way and then move to the more formal definitions.

The first feature that we need to consider in order to model the grounding relation is linked to the fact that grounding amounts to a *dependence* relation of the conclusion on its grounds. This dependence, in its turn, can be conveyed by saying that if the grounds were modified somehow (under certain conditions), then this change would affect the conclusion. In other terms, in a grounding relation, not only does the conclusion logically follow from the grounds, but also the negation of the conclusion needs to logically follow from the negation of some (even all) the grounds (under certain conditions).[17]

The conclusion is thus dependent on its grounds; is dependence all there is to a grounding relation? A glimpse at the explanatory literature is enough to answer negatively to this question. Indeed, any explanatory relation, such as grounding, is asymmetric: there is a direction from what explains to what is explained. The dependence does not provide such a directionality. To see this clearly, one can consider the case of any unique ground, say $F \Vdash A$. In this type of case, dependence boils down to an equivalence between F and A and thus we need an ingredient which establishes why it is the case that F is the ground of A, and not viceversa.[18] According to a long-standing and illustrious philosophical tradition, the required ingredient is complexity: from Aristotle to Bolzano,[19] passing through, amongst others, Arnauld and Nicole,[20] scholars tend to agree that the simplest premisses ground the more complex conclusion, and it would be absurd to go the other way. Moreover, increase in complexity from the grounds to their conclusion should be of a particular type: the formulas by means of which a sentence is grounded should correspond to a decomposition of the sentence itself.[21] Although this insight is clear, deep, and supported by a brilliant tradition, problems arise when we try to formalize it. The first notions that would seem to naturally serve the purpose are logical complexity and associated relation of subformula; however, as has been noticed in several papers,[22] they turn out to be inadequate for an explanatory framework. More precisely, there are two kinds of counterexample which can be evoked to show that complexity and subformula are not adequate for explanation. The first type of example - example (i) - concerns the use of negation. Consider the following sentence: (a) "it is not the case that it rains or it is windy." Suppose one aims at identifying the (total and immediate) reasons why this sentence is true. These typically amount to the sentences (b)"it is not the case that it rains," and (c) "it is not the case that it is windy." Let us formalize (a)-(c), then we get (a) $\neg(p \vee q)$, (b) $\neg p$, (c) $\neg q$. Although we would like

[16] E.g. see Poggiolesi (2016); Poggiolesi and Francez (2021).

[17] E.g. see Jansson (2017); Poggiolesi and Francez (2021).

[18] Note that this is precisely the case for mathematical example mentioned above, where a property of circles is explained by an unique reason, namely a property of points. It turns out that these two properties are equivalent, i.e. the property of the circles logically follows for the property of the points, as the property of points logically follows from the property of circles.

[19] E.g. see de Jong and Betti (2010); Betti (2010); Rumberg (2013).

[20] See Arnauld and Nicole (1993).

[21] See Roski and Rumberg (2016).

[22] See Arana (2009); Kahle and Pulcini (2017).

$\neg p$ and $\neg q$ to be less complex[23] and subformulas of $\neg(p \vee q)$, according to the standard notion of logical complexity and subformula, they are not. Hence this is the first type of counterexample to logical complexity and subformula in an explanatory framework.

Let us move to the second type of counterexample - example (ii). Consider the sentence (a) "for any x, if x is zero or it is the successor of natural number, then it is itself a natural number." Suppose one aims at identifying the (total and immediate) reasons why this sentence is true. These seem to amount to the sentence (b) "for any x, if x is zero, then it is a natural number" and (c)"for any x, if x is the successor of a natural number, then it is a natural number." Let us formalize (a)-(c), then we get (a) $\forall x((Zx \vee SNx) \rightarrow Nx)$, (b) $\forall x(Zx \rightarrow Nx)$, (c) $\forall x(SNx \rightarrow Nx)$. Although we would like (b) and (c) to be logically less complex and subformulas of (a), according to the standard notion of logical complexity and subformula, they are not. Hence this is the second type of counterexample to the use of logical complexity and subformula in an explanatory framework.

Although the notions of logical complexity and subformula are central for many logical results and cornerstones of proof theory, they mainly stand as technical devices which do not necessarily reflect philosophical perspectives. The insight is thus to enrich them both so that they become adequate for grounding and explanation. This will require several definitions that we will introduce formally and then clarify.

Definition 3.1. Let $A \in \mathcal{WF}$, the g-complexity of A, $gcm(A)$, is defined in the following way:

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

In Definition 3.1 we provide a novel way of counting the complexity of a formula that is adequate for an explanatory framework and relies on the work of Poggiolesi (2016), extending it at the first-order level. The main insight behind the notion of g-complexity is that it tracks relationships among the truths expressed by the formulas if they were true. It does that because it is a notion that aims to be apt for explanation and grounding, and both these relations are mainly concerned with truths.[24] Let us now see how g-complexity works. Consider first conjunction, disjunction and quantifiers: in these cases, g-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. Analogously, if Pc expresses a truth, then the truth expressed by $\forall x Px$ is obtained from the previous truths using a single operation. However, this is not so for the case of negation. 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: in other terms, $\neg Pc$ can no longer be seen as constructed from Pc, since if one is true, the other is false. We should rather look at them as two formulas on the same level and this is precisely what the g-complexity does. Analogous reasoning can be applied to the g-complexity of more complex formulas like $A \wedge B$ and

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

[24] E.g. see Correia and Schnieder (2012).

$\neg(A \wedge B)$, or $\forall x Ax$ and $\neg(\forall x Ax)$. 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 the two faces of the same medal, two formulas at the same level and thus having the same g-complexity. Let us finally move to the case of double negation. In this case, the negation counts since $gcm(\neg\neg A) = gcm(A) + 1$. But this is in harmony with what has been said up to now as $\neg\neg A$ and A may both express truths, and thus the former can be seen as constructed from the latter by means of a single operation.

Related to the new notion of g-complexity, we introduce the notion of *converse of a formula*.

Definition 3.2. 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, $(M)^* := \{B^* | B \in M\}$.

Consider a formula A. The converse of A, i.e. A^*, is a formula such that $A \wedge A^*$ forms a contradiction and A and A^* have the same g-complexity.

Now that we have a new notion of g-complexity, we can move to the main ideas behind the new related notion of subformula we aim at proposing, that will be called *g-subformula*. To convey these ideas, consider the example of the formula $B = \exists x(Sx \wedge Tx) \vee \forall x \forall y(Px \rightarrow Qx \wedge Ry)$. The standard (immediate) subformulas of B are $B' = \exists x(Sx \wedge Tx)$ and $B'' = \forall x \forall y(Px \rightarrow Qx \wedge Ry)$. According to the new measure of g-complexity both B' and B'' are still less g-complex than B and thus we still consider them as (immediate) g-subformulas of B. However, they will no longer be the only (immediate) g-subformulas of B as we will enrich the standard notion of subformula by incorporating the following three main ideas.

1. The first idea is linked to the notion of converse. A formula A, and its converse A^*, are now the two faces of a same medal: they concern the same state of affairs and they occupy the same place in the explanatory hierarchy, i.e., they have the same g-complexity. As a result, whenever a formula B' is a g-subformula of a formula B, also its converse will be. Thus, in our example, $(B')^* = \neg\exists x(Sx \wedge Tx)$ and $(B'')^* = \neg\forall x \forall y(Px \rightarrow Qx \wedge Ry)$ will also be (immediate) g-subformulas of B.

2. The second idea is linked to the fact that, standardly, in order to obtain subformulas, we break the formula in question along its main connective. However, in an explanatory framework, where the focus often goes on parts of formulas that do not correspond to the main connective,[25] this operation is restrictive. We thus enrich it in the following way. Consider again the example of the formula $B = \exists x(Sx \wedge Tx) \vee \forall x \forall y(Px \rightarrow Qx \wedge Ry)$. Suppose we want to focus on the part of B that corresponds to $Qx \wedge Ry$. We denote this fact by rewriting B as $D[Qx \wedge Ry]$; in other words, $D[]$ denotes the part of the formula B which corresponds to $\exists x(Sx \wedge Tx) \vee \forall x \forall y(Px \rightarrow$. Having switched the focus on this new part of the formula, we will break it at that point.[26] More precisely, we will have that the

[25] To see this consider the mathematical example of Section 1., or the counterexample concerning the formula $\forall x((Zx \vee SNx) \rightarrow Nx)$ examined above.

[26] A very similar idea, although motivated by different insights, has been put forward by, e.g. Guglielmi and Bruscoli (2009) and their extensive work on *deep inferences*, e.g. http://alessio.guglielmi.name/res/cos/index.html.

g-subformulas of $Qx \wedge Ry$ will remain so even inserted in $D[]$. Hence, $D[Qx]$ and $D[Ry]$ are g-subformulas of $D[Qx \wedge Ry]$, where

$$D[Qx] = \exists x(Sx \wedge Tx) \vee \forall x \forall y(Px \rightarrow Qx)$$

$$D[Ry] = \exists x(Sx \wedge Tx) \vee \forall x \forall y(Px \rightarrow Ry)$$

But also, because of what has been said at point 1, $D[Qx^*]$, $D[Ry^*]$, $(D[Qx])^*$, $(D[Ry])^*$, $(D[Qx^*])^*$ and $(D[Ry^*])^*$ will be g-subformulas of $D[Qx \wedge Ry]$, which is nothing but the formula B.

3. The third and last idea consists in closing the relation of g-subformula under associativity and commutativity of conjunction and disjunction, change of orders of identical quantifiers, and substitution of variables. This means that if B' is a g-subformula of B, then any formula C, which is equivalent to B' by associativity and commutativity of conjunction and disjunction, change of orders of identical quantifiers, and substitution of variables, is also a g-subformula of B. The idea is motivated by noticing that B' and C are interchangeable from an explanatory perspective and thus the g-subformula relation should account for such a fact.[27]

Now that we have clarified the main insights behind the new notion of *g-subformula*, we introduce it in a formal way via the following definitions.

Definition 3.3. The set Co of contexts is inductively defined in the following way:

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

Definition 3.4. For all contexts $F[.]$, and formulas C, we define $F[C]$, a formula in a context, as follows:

- if $F[.] = [.]$, then $F[C] = C$,
- if $F[.] = \neg\neg E[.]$, then $F[C] = \neg\neg E[C]$,
- if $F[.] = G \circ E[.]$, $E[.] \circ G$, $\odot x E[.]$, $\neg E[.]$, then $F[C] = G \circ E[C]$, $E[C] \circ G$, $\odot x E[C]$, $\neg E[C]$, respectively.

Definition 3.5. We define the g-complexity of a context $gcm(F[.]) = gcm(F[Pc])$ for any predicate P and constant c in $\mathcal{L}$.

Definition 3.6. We define the g-complexity of a formula in context, $gcm(F[C])$ as a pair of numbers (m, n) such that $m = gcm(F[.])$ and $n = gcm(C)$. Accordingly, given the formulas in a context $F_1[C_1], ..., F_k[C_k]$ and $G[D]$, such that $gcm(F_1[C_1]) = (m, n_1)$, ..., $gcm(F_k[C_k]) = (m, n_k)$ and $gcm(G[D]) = (m, n)$, where $n = n_1 + ... + n_k + 1$, $F_1[C_1], ..., F_k[C_k]$ will be said to be *immediately less g-complex* than $G[D]$.

[27] This feature is known in the literature as *ground-theoretic* or *factual* equivalence, e.g. see Correia (2016, 2017).

Definition 3.7. Given a formula A of $\mathcal{L}$, we say that A is *FOL-equiv* to B if, and only if, A can be obtained from B by associativity and commutativity of conjunction and disjunction, substitution of variables, and change of orders of identical quantifiers.

Definition 3.8. Given a context $F[.]$ of $\mathcal{L}$, we say that $F[.]$ is *FOL-equiv* to $G[.]$ if, and only if, for any predicate P and any constant $c \in \mathcal{L}$, $F[Pc]$ is FOL-equiv to $G[Pc]$.

Definition 3.9. For any pair of formulas of $\mathcal{L}$ A and B, we say that $A \cong B$ if, and only if, A is FOL-equiv to B or A is FOL-equiv to B^*.

Definition 3.10. For any pair of contexts of $\mathcal{L}$ $F[.]$ and $G[.]$, we say that $F[.] \cong G[.]$ if, and only if, for any predicate P and any constant c in $\mathcal{L}$, $F[Pc]$ is FOL-equiv to $G[Pc]$ or $F[Pc]$ is FOL-equiv to $(G[Pc])^*$.

Definition 3.11. For any pair of multisets of $\mathcal{L}$ M and N, such that $M = \{A[C_1], ..., A_n[C_n]\}$ and $N = \{B_1[D_1], ..., B_n[D_n]\}$, we say that $M \cong N$, if, and only if, $A_1 \cong B_1$, ..., $A_n \cong B_n$ and $C_1 \cong D_1$, ..., $C_n \cong D_n$.

Definition 3.12. For any pair of formulas in contexts of $\mathcal{L}$ $F[B]$ and $G[A]$, we say that $F[B]$ is a g-subformula of $G[A]$ if, and only if, $F[.] \cong G[.]$, and:

- $A \cong B$,
- $A \cong \neg\neg C$ and B is a g-subformula of C,
- $A \cong C \circ D$ and B is a g-subformula of C or B is a g-subformula of D,
- $A \cong \odot x C$ and B is a g-subformula of $C(t/x)$ for all t free for x in C.

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

Definition 3.13. M is a multiset of *distinguished immediate g-subformulas* of $G[A]$, if, and only if:

- $M \cong \{G[B]\}$ and $A \cong \neg\neg B$,
- $M \cong \{G[B], G[D]\}$ and $A \cong (B \circ D)$,
- $M \cong \{G[\odot x B\}$ and $A \cong B(t/x)$, for all t free for x in B.

Note that the distinguished immediate g-subformulas of $G[A]$ are always immediately less g-complex than $G[A]$ according to Definition 3.6, so that the notion of g-complexity and g-subformula go hand in hand.

We now have almost all the elements required to properly define the relation of formal grounding. We only need to introduce the last two, which will have a central role for what follows. The first stems from the observation that contexts, as well as formulas in contexts, can be assigned a related (positive or negative) polarity, which is defined in a standard way as follows.[28]

Definition 3.14. We define positive $\mathcal{P}$ and negative (formula-)polarities $\mathcal{N}$ simultaneously by an inductive definition given by the three clauses (i)-(iii) below.

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

[28] See Troelstra and Schwichtenberg (1996).

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

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

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

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

We can now introduce the notion of *scope of a context* (and the inverse scope), which, given a context $F[.]$, corresponds to the list of consecutive quantifiers in F (selected also according to their polarities) B is in the scope of.

Definition 3.15. If $F[.]$ is a context, the *scope of a context*, $SC(F)$ and the inverse scope $SC^{inv}(F)$ are defined inductively in the following way:

- if $F[.]= [.]$ or $F[.] \neq \overbrace{\neg...\neg}^{2n}[.]$ for $n \geqslant 0$ then $SC(F) = SC^{inv}(F) = \varnothing$,
- if $F[.]= G \circ E[.]$ or $E[.] \circ G$, then $SC(F) = SC(E)$ and $SC^{inv}(F) = SC^{inv}(E)$,
- if $F[.]= \forall x E[.]$, then $SC(F) = \forall x.(SC(E))$ and $SC^{inv}(F) = \exists x.SC^{inv}(E)$
- if $F[.]= \exists x E[.]$, then $SC(F) = \exists x.(SC(E))$ and $SC^{inv}(F) = \forall x.SC^{inv}(E)$
- if $F[.]= \neg E[.]$, then $SC(F) = SC^{inv}(E)$ and $SC^{inv}(F) = SC(E)$.

Definition 3.16. For any finite multisets of $\mathcal{CF}$ $M = \{A_1[D_1], ..., A_m[D_m]\}$ and $N = \{A'_1[C_1], ..., A'_n[C_n]\}$(which could be empty), and for any $\mathcal{CF}$ $F[B]$, under the condition that N^*, M is a *total and immediate formal ground* of $F[B]$, in symbols $N \mid M \Vdash F[B]$, if, and only if, for any E such that $SC(E) = SC(F)$ and $E \in \mathcal{P}$ if, and only if, $F \in \mathcal{P}$, we have:

1. $E[D_1], ..., E[D_m] \models E[B]$,
2. for some non empty (possibly non proper) submultiset M' of M, such that $M' = \{A_{k1}[D_{k1}], ..., A_{kr}[D_{kr}]\}$, we have that $(E[C_1])^*, ..., (E[C_n])^*, (E[D_{k1}])^*, ..., (E[D_{kr}])^*, M^-/E \models (E[B])^*$.
3. $N \cup M$ is a multiset of immediate and distinguished g-subformulas of $F[B]$.

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

Definition 3.16 represents the formal counterpart of the features discussed in this section. Conditions 1. and 2. are meant to capture grounding as a dependence relation. Obviously this dependence holds amongst the formulas (in contexts) at issue, independently from the contexts these formulas belong to. For this reason, universal quantification over any context $E[.]$, whose scope and polarity are the same as that of the formula to be grounded, needs to be added. Condition 3. amounts to the directionality or asymmetry of the grounding relation: this is conveyed via the new notion of g-subformula.

Let us evaluate some grounding principles which emerge from this definition. Consider the formula $\neg(p \vee q)$ that we have discussed in the example (i) above; as we have already said, standardly $\neg p, \neg q$ are taken to be the total and immediate grounds for this formula. Definition 3.16 confirms this intuition: indeed $\neg(p \vee q)$ is a classical logical consequence of

$\neg p$ and $\neg q$. However, it is also the case that if we modify the grounds 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 immediate and distinguished g-subformulas of $\neg(p \vee q)$.

Let us now move to the formula $\forall x((Zx \vee SNx) \rightarrow Nx)$ that we have discussed in the example (ii) above. In particular we have said that it is intuitive to take the formulas $\forall x(Zx \rightarrow Nx)$ and $\forall x(SNx \rightarrow Nx)$ to be its total and immediate ground. Definition 3.16 confirms this intuition: indeed 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 $E[SNx]$. However, it is also the case, that if we modify the grounds, 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 immediate and distinguished g-subformulas of $\forall x((Zx \vee SNx) \rightarrow Nx)$.

Finally, consider the sentence "for any x, if x is a natural number, then it is an odd or an even number," that we formalize with $\forall x(Nx \rightarrow Ex \vee Ox)$. 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 g-subformulas of $\forall x(Nx \rightarrow Ex \vee Ox)$, it would be rather weird to think of them as its grounds, if only because they are false. Our model confirms this intuition in that it can be easily checked that condition 2. of Definition 3.16 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 $\forall x(Nx \rightarrow Ox)$. Hence, even for the negative cases, our model seems to go hand in hand with our intuitions.

We can extend the definition of total and immediate formal grounding to total and mediate formal grounding in the following way.

Definition 3.17. For any multisets of $\mathcal{CF}$ M and N (which could be empty), and for any $\mathcal{CF}$ B, under the condition that N^*, M is a *total and mediate formal ground* of B, $N \mid M \Vdash_m B$, if, an only if:

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

4 Formal explanation

In this section we move to consider the relation of formal explanation, which we see as a strenghtening of the classical logical derivability relation. Hence, we will construct formal explanations as we construct derivations, namely by firstly introducing rules and then by formally defining the notion via the rules. Note that we want formal explanations to go hand in hand with formal grounding; namely, we want a notion of formal explanation such that there exists a formal explanation of A from M, under conditions N^* if, and only if, M are the reasons why A is true, under conditions N^*. Not only will we shape our notion of formal explanation around this desideratum, in Section 5 we will prove that the desideratum has been met.

In order to introduce 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 2) is equivalent to the classical sequent calculus for first-order logic **Gcl**. The notion of sequent, its interpretation, and

Figure 2: The sequent calculus $\mathbf{Gcl}^{+}$.

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

$$\frac{M \Rightarrow N, A}{\neg A, M \Rightarrow N}\ \neg L \qquad \frac{A, M \Rightarrow N}{M \Rightarrow N, \neg A}\ \neg R \qquad \frac{A, B, M \Rightarrow N}{A \wedge B, M \Rightarrow N}\ \wedge L \qquad \frac{M \Rightarrow N, A \quad M \Rightarrow N, B}{M \Rightarrow N, A \wedge B}\ \wedge R$$

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

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

the interpretation of inferential rules are standard (e.g. see Troelstra and Schwichtenberg (1996)). We call *c-sequent* a sequent that only contains closed formulas.

Definition 4.1. Explanatory rules will have one of the following three forms, together with their associated interpretations:

- $\dfrac{\dfrac{M' \Rightarrow N'}{M \Rightarrow N}}{}$ $:= \bigwedge M' \rightarrow \bigvee N' \Vdash \bigwedge M \rightarrow \bigvee N$
- $\dfrac{\dfrac{M' \Rightarrow N' \quad M'' \Rightarrow N''}{M \Rightarrow N}}{}$ $:= \bigwedge M' \rightarrow \bigvee N', \bigwedge M'' \rightarrow \bigvee N'' \Vdash \bigwedge M \rightarrow \bigvee N$
- $\dfrac{\dfrac{M' \Rightarrow N' \mid M'' \Rightarrow N''}{M \Rightarrow N}}{}$ $:= \bigwedge M' \rightarrow \bigvee N' \mid \bigwedge M'' \rightarrow \bigvee N'' \Vdash \bigwedge M \rightarrow \bigvee N$

Definition 4.2. We say that a context $F[.]$ has

a *positive universal scope* (PUS) if $F \in \mathcal{P}$ and $SC(F) = \varnothing$ or $SC(F) = \forall x_1, ...\forall x_n$,
a *negative universal scope* (NUS) if $F \in \mathcal{N}$ and $SC(F) = \varnothing$ or $SC(F) = \forall x_1, ...\forall x_n$,
a *positive existential scope* (PES) if $F \in \mathcal{P}$ and $SC(F) = \varnothing$ or $SC(F) = \exists x_1, ...\exists x_n$,
a *negative existential scope* (NES) if $F \in \mathcal{N}$ and $SC(F) = \varnothing$ or $SC(F) = \exists x_1, ...\exists x_n$.

Definition 4.3. We assume the application of explanatory propositional rules[30] to obey the following restrictions:

- rule $\circ_1$ can be applied on a formula of the form $F[B \circ C]$ if: $\begin{cases} F \text{ has PUS and } \circ = \wedge, \text{or} \\ F \text{ has NUS and } \circ = \vee, \text{or} \\ F \text{ has PES and } \circ = \vee, \text{or} \\ F \text{ has NES and } \circ = \wedge. \end{cases}$
- rule $\circ_2$ can be applied on a formula of the form $F[B \circ C]$ if: $\begin{cases} F \text{ has PES and } \circ = \vee, \text{or} \\ F \text{ has NES and } \circ = \wedge. \end{cases}$
- rule $\neg\circ_1$ can be applied on a formula of the form $F[\neg(B \circ C)]$ if: $\begin{cases} F \text{ has PUS and } \circ = \vee, \text{or} \\ F \text{ has NUS and } \circ = \wedge, \text{or} \\ F \text{ has PES and } \circ = \wedge, \text{or} \\ F \text{ has NES and } \circ = \vee. \end{cases}$

[30] Reading the rules bottom-up.

Figure 3: Explanatory propositional rules.

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

$$\frac{M \Rightarrow N, F[B] \quad M \Rightarrow N, F[C]}{M \Rightarrow N, F[B \circ C]}\ \circ_1 \qquad \frac{M \Rightarrow N, F[B_j] \mid \quad M \Rightarrow N, F[B_i]}{M \Rightarrow N, F[B_1 \circ B_2]}\ \circ_2$$

$$\frac{M \Rightarrow N, F[B^*] \quad M \Rightarrow N, F[C^*]}{M \Rightarrow N, F[\neg(B \circ C)]}\ \neg\circ_1 \qquad \frac{M \Rightarrow N, F[B_j^*] \mid \quad M \Rightarrow N, F[B_i^*]}{M \Rightarrow N, F[\neg(B_1 \circ B_2)]}\ \neg\circ_2$$

where both i and $j = \{1, 2\}$ and $j \neq i$.
We assume explanatory propositional rules not to distinguish between formulas which are FOL-equiv,[29] and to only apply to c-sequents. Their application is conditioned by Definition 4.3.

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

Explanatory rules are such that not only is the conclusion derivable from the premise(s), but also the premisses are the total and immediate reasons why the conclusion is true. Differently from the inferential rules which only operate on the main connective of a formula, explanatory rules can modify a formula from the inside, i.e., along connectives different from the main ones. In explanatory rules it is thus very important to check the polarity of the formula in a context, as well as the type of scope of the context, one is dealing with. Let us provide some examples of application of explanatory propositional rules (see Figure 3). The first one is the following, which corresponds to our previous case (i):

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

Thanks to the rule $\neg\circ_1$, we can explain, totally and immediately, the formula $\neg(p \vee q)$ by the formulas $\neg p$ and $\neg q$, which are its ground. The rule is applicable because $\neg(p \vee q)$ lies in the scope of no quantifier. Let us now move to another example, which corresponds to our previous case (ii):

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

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 $\circ_1$ can be applied on a formula of the form $Zx \vee SNx$ in a context F, since F has NUS (see Definition 4.3). Hence the rule matches our previous intuitions. Finally, one can easily check that no rule provides the grounds of the formula $F[Ex \vee Ox] = \forall x(Nx \rightarrow Ex \vee Ox)$. Indeed no rule is applicable on a formula in a context whose main connective is $\vee$, and such that the context F has PUS (see again Definition 4.3).

As for explanatory rules for quantifiers (see Figure 4), we had a choice: we could either go towards infinitary rules, or remain in the finite. The choice of infinitary rules, which is

Figure 4: Explanatory first-order rules.

$$\frac{M \Rightarrow N, F[By]}{M \Rightarrow N, F[\odot x.Bx]}\ \odot_1 \qquad \frac{M \Rightarrow N, F[\odot x.Bx], F[Bt]}{M \Rightarrow N, F[\odot x.Bx], F[\odot x.Bx]}\ \odot_2$$

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

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

We assume explanatory first-order rules not to distinguish between formulas which are FOL-equiv. Their application is conditioned by Definition 4.4.

the most followed in the literature,[31] involves an enriched language - with one constant for each element of the domain - thus an enriched sequent calculus as well. Although infinitary rules provide a reasonable intuition of how explanation in logic might work, their main disadvantage is that they are proof-theoretically unsatisfactory. The finitary rules, on the other hand, require a change in the formulas involved in explanations - we move from closed formulas to open formulas - but they are proof-theoretically satisfactory. In this paper, following the work of Genco et al. (2021), we choose to deal with finitary rules (see Figure 4). Roughly speaking, the rule for the universal quantifier explains this quantifier by using the *eigenvariable*,[32] i.e. it explains why any object x has a property B via the fact that if one picks a random object y, y has the property B. 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 grounds of its conclusion, the existential itself needs to be repeated in the premisses and doubled in the conclusion.[33] Thus, whilst the rule for the universal quantifier is adequate both at the proof-theoretical and intuitive level, the rule for the existential quantifier although proof-theoretically acceptable, is less satisfactory as long as intuitions are concerned.

Definition 4.4. We assume the application of explanatory first-order rules[34] to obey the following restrictions:

- rule $\odot 1$ can be applied on a formula of the form $F[\odot x.Bx]$ if: $\begin{cases} F \text{ has } PUS \text{ and } \circ{=}\forall, \text{or} \\ F \text{ has } NUS \text{ and } \circ{=}\exists, \text{or} \\ F \text{ has } PES \text{ and } \circ{=}\exists, \text{or} \\ F \text{ has } NES \text{ and } \circ{=}\forall. \end{cases}$
- rule $\odot_2$ can be applied on a formula of the form $F[\odot x.Bx]$ if $F \in \mathcal{P}$ and $\odot = \exists$, or $F \in \mathcal{N}$ and $\odot = \forall$.
- rule $\neg\odot_1$ can be applied on a formula of the form $F[\neg(\odot x.Bx)]$ if: $\begin{cases} F \text{ has } PUS \text{ and } \circ{=}\exists, \text{or} \\ F \text{ has } NUS \text{ and } \circ{=}\forall, \text{or} \\ F \text{ has } PES \text{ and } \circ{=}\forall, \text{or} \\ F \text{ has } NES \text{ and } \circ{=}\exists. \end{cases}$

[31] E.g. see Correia (2017); Fine (2012).
[32] See Troelstra and Schwichtenberg (1996).
[33] 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).
[34] Reading the rules bottom-up.

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

We will call $\mathbf{Gcl}^E$ the sequent calculus composed by the rules of Figures 2 and 3, whilst we will call $\mathbf{Gcl}^{EQ}$ the sequent calculus composed by the rules of Figures 2, 3 and 4. In what follows, for us to keep on working with closed formulas, we will mainly deal with the calculus $\mathbf{Gcl}^E$, leaving results concerning $\mathbf{Gcl}^{EQ}$ for future research.

Definition 4.5. A *mixed derivation* in $\mathbf{Gcl}^E$ is a finite (upwardgrowing) tree with a single root. The nodes of the tree are labelled by sequents or sequents with a bar and the top nodes are labelled by initial sequents. For each non-terminal node, its label is connected with the labels of the immediate predecessor nodes according with one of the logical rules or one of the explanatory rules, or the rule *cw*. The root of the tree is the conclusion of the whole derivation and its label is a theorem of the sequent calculus, in symbol $\vdash^{\star}_{\mathbf{Gcl}^E} M \Rightarrow N$. A *derivation* in $\mathbf{Gcl}^E$ is a mixed derivation where only logical rules have been applied; we denote it as usual with the symbol $\vdash_{\mathbf{Gcl}^E}$.

In the calculus $\mathbf{Gcl}^E$ it is thus possible to construct standard derivations, but also derivations which contain explanatory steps that we call *mixed derivations*.

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

Definition 4.6. For any multisets of c-sequents S and S' (which might be empty), and for any c-sequent $M \Rightarrow N$, we say that under the condition $(S')^*$, there exists a *total and immediate formal explanation* from S to $M \Rightarrow N$, in symbols $S' \mid S \Vdash M \Rightarrow N$ if, and only if, one of the explanatory rules of Figure 3 links S', S and $M \Rightarrow N$.

Definition 4.7. For any multisets of c-sequents S and S' (which might be empty), and for any c-sequent $M \Rightarrow N$, we say that under the condition $(S')^*$, there exists a *total and mediate formal explanation* from S to $M \Rightarrow N$, in symbols $S' \mid S \Vdash_m M \Rightarrow N$ if, and only if:

- $S' \mid S \Vdash M \Rightarrow N$,
- $S'' \mid S''' \Vdash M' \Rightarrow N'$ *and* $S'''' \mid S'''''$, $M' \Rightarrow N' \Vdash_m M \Rightarrow N$, and $S'' \cup S'''' = S'$ and $S''' \cup S''''' = S$.

5 Soundness and completeness

We use this section to prove soundness and completeness between the notions of formal grounding and formal explanation. We will start by soundness, and for that, we will first prove some preliminary lemmas.

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

when $A \in \mathcal{P}$: *when* $A \in \mathcal{N}$:

$$\frac{A[B_i], M \Rightarrow N}{A[B_1 \wedge B_2], M \Rightarrow N}\ {\wedge}1 \qquad \frac{M \Rightarrow N, A[B_i]}{M \Rightarrow N, A[B_1 \vee B_2]}\ {\vee}1 \qquad \frac{M \Rightarrow N, A[B_i]}{M \Rightarrow N, A[B_1 \wedge B_2]}\ {\wedge}2 \qquad \frac{A[B_i], M \Rightarrow N}{A[B_1 \vee B_2], M \Rightarrow N}\ {\vee}2$$

$$\frac{A[B_i^*], M \Rightarrow N}{A[\neg(B_1 \vee B_2)], M \Rightarrow N}\ \neg{\vee}1 \qquad \frac{M \Rightarrow N, A[B_i^*]}{M \Rightarrow N, A[\neg(B_1 \wedge B_2)]}\ \neg{\wedge}1 \qquad \frac{M \Rightarrow N, A[B_i^*]}{M \Rightarrow N, A[\neg(B_1 \vee B_2)]}\ \neg{\wedge}2 \qquad \frac{A[B_i^*], M \Rightarrow N}{A[\neg(B_1 \wedge B_2)], M \Rightarrow N}\ \neg{\vee}2$$

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 the context $A[.]$, and subinduction on the height of the derivation of the premise of the rule. The admissibility of any other rule can be proved analogously.

We distinguish cases according to the form of $A[.]$. If $A[.] = [.]$, then from the premise $B_i, M \Rightarrow N$ we obtain the desired result thanks to the rule $\wedge L$. As for the rule $\wedge 2$, since B_i has a negative polarity in $A[.]$, it can be thought of as $\neg(B_i)$. Thus we first apply the inverse of the rule $\neg R$[35] obtaining $B_i, M \Rightarrow N$. We then apply the rules $\wedge L$ and $\neg R$ to get the desired result.

If $A[.] \neq [.]$, then we distinguish cases according to the last applied rule $\mathcal{R}$ on $A[B_i], M \Rightarrow N$ and on $M \Rightarrow N, A[B_i]$. • 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. • A rule $\mathcal{R}$ has been applied on $A[B_i]$ in the sequent $A[B_i], M \Rightarrow N$ (the case where $\mathcal{R}$ has been applied on $A[B_i]$ in the sequent $M \Rightarrow N, A[B_i]$ is analogous). We distinguish the following subcases according to the form of A.

$$\frac{E, F[B_i], M \Rightarrow N}{E \wedge F[B_i], M \Rightarrow N} \rightsquigarrow^{36} \frac{E, F[B_1 \wedge B_2], M \Rightarrow N}{E \wedge F[B_1 \wedge B_2], M \Rightarrow N}$$

$$\frac{E, M \Rightarrow N \quad F[B_i], M \Rightarrow N}{E \vee F[B_i], M \Rightarrow N} \rightsquigarrow \frac{E, M \Rightarrow N \quad F[B_1 \wedge B_2], M \Rightarrow N}{E \vee F[B_1 \wedge B_2], M \Rightarrow N}$$

$$\frac{\forall x F[B_i], F[B_i], M \Rightarrow N}{\forall x F[B_i], M \Rightarrow N} \rightsquigarrow \frac{\dfrac{\forall x F[B_1 \wedge B_2], F[B_i], M \Rightarrow N}{\forall x F[B_1 \wedge B_2], F[B_1 \wedge B_2], M \Rightarrow N}\ i.h.}{\forall x F[B_1 \wedge B_2], M \Rightarrow N}$$

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

□

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

$$A[B] \Vdash A[\neg\neg B]$$

[35] All logical rules are invertible in **Gcl**, see Troelstra and Schwichtenberg (1996)).

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

[37] The case where $A[B_i]$ is of the form $\neg\neg F[B_i]$ is clearly analogous.

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

Definition 5.3. Given $A, B, C \in \mathcal{CF}$, by

$A, B \div C$ we denote $A, B \models C$ and $A^*, B^* \models C^*$.

$A \mid B \div C$ we denote $B \models C$ and $A^*, B^* \models C^*$.

$< A > B \div C$ we denote $B^* \models C^*$ and $A, B \models C$.

Lemma 5.4. *For any $A, B, C \in \mathcal{CF}$:*

$$A \mid B \div C \text{ if, and only if, } < A^* > B^* \div C^*$$

Proof. Straightforward. □

Lemma 5.5. *For any context $D[.]$ that has PUS (see Definition 5.11) and for any formula $G, G', C \in \mathcal{CF}$, such that $C \in \{G \wedge G', \neg(G \vee G')\}$, then it holds that:*

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

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

For any context $D[.]$ that has NES (see Definition 5.11) and for any formula $G, G', C \in \mathcal{CF}$, such that $C \in \{G \wedge G', \neg(G \vee G')\}$, then it holds that:

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

(d) if $< G > G' \div C$, then $D[G] \mid D[G'] \div D[C]$.

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

(i). It is straightforward.

(ii). Suppose $G, G' \div C$ (the other option is to have $G^*, G'^* \div C$. This can be treated analogously). By i.h., one obtains $F[G], F[G'] \div F[C]$. In order to get the desired result, we exploit the sequent calculus **Gcl** in the following way:[40]

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

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

[39] The case $D = F[.] \vee E$ 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**.

From $E \wedge F[G], E \wedge F[G'] \vdash E \wedge F[C]$ by completeness of **Gcl**, one gets $E \wedge F[G], E \wedge F[G'] \models E \wedge F[C]$. From $E \wedge F[C] \vdash E \wedge F[G] \vee E \wedge F[G']$ by completeness of **Gcl**, and the symbol of converse (see Definition 3.2), one gets $(E \wedge F[G])^*, (E \wedge F[G'])^* \models (E \wedge F[C])^*$. Thus we have $E \wedge F[G], E \wedge F[G'] \doteqdot E \wedge F[C]$.

(iii). Analogously to (ii).

(iv) In this case we further distinguish sub-cases according to the form of C. We thus have (iva) $C = G \wedge G'$, and (ivb) $C = \neg(G \vee G')$.

(iva). By i.h., one obtains $F[G], F[G'] \doteqdot F[G \wedge G']$. 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]

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

From $\forall xF[Gx], \forall xF[G'x] \vdash \forall xF[Gx \wedge G'x]$ by completness of **Gcl** one gets $\forall xF[Gx], \forall xF[G'x] \models \forall xF[Gx \wedge G'x]$. From $\forall xF[Gx \wedge G'x] \vdash \forall xF[Gx] \vee \forall xF[G'x]$ by completeness of **Gcl**, and the symbol of converse (see Definition 3.2), one gets $(\forall xF[Gx])^*, (\forall xF[G'x])^* \models (\forall xF[Gx \wedge G'x])^*$. Thus we have $\forall xF[Gx], \forall xF[G'x] \doteqdot \forall xF[Gx \wedge G'x]$.

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

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

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

□

Lemma 5.6. *For any context $D[.]$ that has PES (see Definition 5.11) and for any formula $G, G', C \in \mathcal{CF}$, such that $C \in \{G \vee G', \neg(G \wedge G')\}$, then it holds that:*

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

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

For any context $D[.]$ that has NUS (see Definition 5.11) and for any formula $G, G', C \in \mathcal{CF}$, such that $C \in \{G \vee G', \neg(G \wedge G')\}$, then it holds that:

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

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

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

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

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 3. 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 $D[B \wedge C]$ such that D has PUS. Clearly, it holds that $B, C \doteqdot B \wedge C$. But, then by Lemma 5.5, we have $\bigwedge M \rightarrow \bigvee N \vee D[B], \bigwedge M \rightarrow \bigvee N \vee D[C] \doteqdot \bigwedge M \rightarrow \bigvee N \vee D[B \wedge C]$, where the context $\bigwedge M \rightarrow \bigvee N \vee D[.]$ has PUS. Actually for Lemma 5.5 again, we have that, for any context $E[.]$ that has PUS, it holds that $E[B], E[C] \doteqdot E[B \wedge C]$. Finally, $\{\bigwedge M \rightarrow \bigvee N \vee D[B], \bigwedge M \rightarrow \bigvee N \vee D[C]\}$ is a multiset of immediate and distinguished g-subformulas of $\bigwedge M \rightarrow \bigvee N \vee D[B \wedge C]$ (also thinking of FOL-equivalent formulas). Hence we have the desired result.

Consider the rule $\circ_1$ applied on a formula of the form $D[B \wedge C]$ such that D has NES. 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 $D[B \vee C]$ such that D has PES. 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 $D[B \vee C]$ such that D has NUS. Then the reasoning is the same as above, except that one needs to use Lemma 5.6.

□

Definition 5.8. For any context $D[.]$, we define the related quantifiers-only-context $Qo(D)[.]$, in the following way:

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

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

Lemma 5.9. *Let $Qo(D)[.]$ be the quantifiers-only-context related to $D[.]$, then:*

$$SC(Qo(D)) = SC(D)$$

Proof. Straightforward from Definition 5.8. □

Lemma 5.10. *For any multisets of $\mathcal{CF}$ M and N (which could be empty), and for any $\mathcal{CF}$ $F[B]$,*

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

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

Proof. By Definition 3.16. □

Definition 5.11. For any quantifier-only-context $Qo(F)[.]$, we say that $Qo(F)[.]$ is:

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

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

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

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

Lemma 5.12. *For any multisets of $\mathcal{CF}$ M and N (which could be empty), and for any $\mathcal{CF}$ $A[C]$,*

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

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 $Qo(F)[.]$ and B.

[-] $Qo(F)[.]$ might be such that: (i) it is a positive universal; (ii) it is a positive existential; (iii) it is a negative universal; (iv) it is a negative existential; (v) $Qo(F)[.]= SC(F)[.]$, where $SC(F)$ 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) $Qo(F)[.]= SC(F)[.]^*$, where $SC(F)$ 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 $\exists x_1, ..., \exists x_n$.

[-] B can be of the following form: (a) $\neg\neg C$; (b) $C \wedge D$; (c) $C \vee D$; (d) $\neg(C \wedge D)$; (d) $\neg(C \vee D)$; (e) $\forall xCx$; (f) $\neg\forall xCx$; (g) $\exists xCx$; (h) $\neg\exists xCx$.

We check in detail the combination 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 $Qo(F)[C] \Vdash Qo(F)[\neg\neg C]$; at the syntactic level the explanatory rule $\neg\neg$ gives us what desired, namely $\Rightarrow F[C] \Vdash \Rightarrow F[\neg\neg C]$.

2. We combine (i) with (b). We have that $Qo(F)[C], Qo(F)[D] \Vdash Qo(F)[C \wedge D]$. At the syntactic level, thanks to the explanatory rule $\circ_1$, we get what desired, namely $\Rightarrow F[C], \Rightarrow F[D] \Vdash \Rightarrow F[C \wedge D]$.

3. We combine (iv) with (b). We have that $Qo(F)[C], Qo(F)[D] \Vdash Qo(F)[C \wedge D]$, $Qo(F)[C] \mid Qo(F)[D] \Vdash Qo(F)[C \wedge D]$ and $Qo(F)[C] \mid Qo(F)[D] \Vdash Qo(F)[C \wedge D]$. At the syntactic level, thanks to the explanatory rules $\circ_1$, $\circ_2$, we get what desired, namely $\Rightarrow F[C], \Rightarrow F[D] \Vdash \Rightarrow F[C \wedge D]$, $\Rightarrow F[C] \mid\Rightarrow F[D] \Vdash \Rightarrow F[C \wedge D]$ and $\Rightarrow F[C] \mid\Rightarrow F[D] \Vdash\Rightarrow F[C \wedge D]$.

4. It is straightforward to check that the combination of (ii), (iii), (v), and (vi) with (b) does not give rise to any grounding principle.

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

□

Theorem 5.13 (Completeness)**.** *For any multisets of closed formulas N, N' (possibly empty), and formula $F[B]$,*

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

Proof. From Lemmas 5.10 and 5.12.

□

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

In this paper 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 $\Vdash$). As for standard derivations, $\mathbf{Gcl}^E$ is equivalent to **Gcl** and it keeps the same properties as **Gcl**.

Lemma 6.1. *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 6.2. *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 for explanatory rules, not only have we shown in the previous section that their premise(s) represent the total and immediate ground of their conclusion, but also we need to show that they do not allow us to derive any new formula, i.e. explanatory rules serve to build *derivations with an explanatory power*, not to prove new theorems. To get this result, we show that any explanatory step from some grounds to their conclusion can also be performed by several applications of the standard inferential rules.

Lemma 6.3. *For any sequent $M \Rightarrow N$, and for any mixed derivation d of $M \Rightarrow N$, namely $d \vdash^{\star} M \Rightarrow N$ which contains only one application of an explanatory rule, one can construct a derivation d' with the same end-sequent, namely $d' \vdash M \Rightarrow N$.*

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 $F[B]$. We apply on the context F as many rules $\overline{R}$ as necessary to unfold the context itself and reach the formula B.[42] Once arrived to B, 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 F. We then apply the standard logic rules

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

to get from B, or any formula FOL-equivalent to B, to the desired conclusion, and then we also use the logical rules to reconstruct the context F. Here is a simple example of the procedure. Consider the following instance of the explanatory rule $\neg\neg$:

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

We obtain the desired result in the following way:

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

As for the rules of the second group, namely those explanatory rules with conditions, one needs to consider the mixed derivation d, which will necessarily contain an application of the rule cw. We substitute the derivation d with a derivation d' with no application of the rule cw. Then we continue the procedure as above. □

Proposition 6.4. *For any sequent $M \Rightarrow N$, and for any mixed derivation d of $M \Rightarrow N$, namely $d \vdash^{\star} M \Rightarrow N$, one can construct a derivation d' with the same end-sequent, namely $d' \vdash M \Rightarrow N$.*

Proof. Considering the derivation d from top, by several applications of the previous Lemma 6.3. □

7 Conclusions

Although the contemporary logical literature abounds with formalizations of key notions, such as knowledge, belief, or time, the equally central notion of explanation has never been given any formal treatment. The main aim of this paper has been to fill this gap and thus to develop a logical theory of the notion of (conceptual) explanation and related notion of grounding. 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 of future research, such as the formalization of the notion of explanation in logics other than classical logic, the applications of formal explanations to related fields such as explainable AI, finally the investigation of the value of explanatory rules in proof-theoretic semantics.

References

Arana, A. (2009). On formally measuring and eliminating extraneous notions in proofs. *Philsophia Mathematica*, 17:189–207.

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

Arnauld, A. and Nicole, P. (1993). *La logique ou l'art de penser*. Vrin, Paris.

Bareinboim, E., Correa, D., Ibeling, D., and Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In Geffner, H., Dechter, R., and Halpern, J., editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. ACM, New York.

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

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

Carrara, M. and De Florio, C. (2020). Identity criteria: an epistemic path to conceptual grounding. *Synthese*, 197:3151–3169.

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

Correia, F. (2017). An impure logic of representational grounding. *Journal of Philosophical Logic*, 46:506–538.

Correia, F. and Schnieder, B. (2012). Grounding: An opinionated introduction. In Correia, F. and Schnieder, B., editors, *Metaphysical grounding*, pages 1–36. Cambridge University Press, Cambridge.

de Jong, W. R. and Betti, A. (2010). The Classical Model of Science I: A Millennia-Old Model of Scientific Rationality. *Synthese*, 174:180–210.

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.

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.

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.

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

Kahle, R. and Pulcini, G. (2017). Towards an operational view of purity. *The Logica Yearbook*, College Publications:56–79.

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.

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

Poggiolesi, F. (2016). 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. 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.

Reutlinger, A. and Saatsi, J. e. (2018). *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*. Oxford University Press, Oxford.

Roski, S. and Rumberg, A. (2016). Simplicity and economy in bolzanoÕs theory of grounding. *Journal of the History of Philosophy*, 54:469–496.

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

Salmon, W. (1989). *Four Decades of Scientific Explanation*. University of Pittsburgh Press, Pittsburgh.

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

Schnieder, B. (2006). A certain kind of trinity: Dependence, substance, and explanation. *Philosophical Studies*, 129:393–419.

Schurz, G. (1999). Explanation as unification. *Synthese*, 120:95–114.

Scriven, M. (1971). The logic of cause. *Theory and Decision*, 2:49–66.

Smithson, R. (2020). Metaphysical and conceptual grounding. *Erkenntnis*, 85:1501–1525.

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.