# Contents

# Introduction

> *The mathematician's patterns, like the painter's or the poet's must be beautiful; the ideas like the colours or the words, must fit together in a harmonious way. Beauty is the first test: there is no permanent place in the world for ugly mathematics.*
>
> Godfried Hardy, *A Mathematician's Apology*

Two broad observations lie at the basis of this dissertation, that finds itself at the intersection between philosophy, mathematics and proof theory. The first one is that mathematicians often prove a theorem in many different ways, even if they already believe in its validity. A theorem that illustrates the pursuit for differing proofs well is Pythagoras's Theorem, having been re-proved many times in its long history — yet even as recently as 2022, a new proof was discovered by New Orleans high school students Ne'Kiya Jackson and Calcea Johnson, which received widespread attention[1], and surely more are to come. Hence, even though Pythagoras's Theorem is one of the most well-established mathematical truths, new proofs are considered to hold value for various reasons. By their use of trigonometric means, Jackson and Johnson's proof can be considered "to demonstrate the power of different methodologies", one of the reasons named by Dawson (2006) to re-prove theorems. Other reasons include improving previous arguments that suffer from perceived gaps or deficiencies; using simpler reasoning; extending or generalizing results; and so on. A specific strand of such motivations focuses on so-called 'ideals' of proof, which are particular conceptual properties that 'good' mathematical proofs are considered to possess (such as simplicity, or beauty, as the above epigraph illustrates). One main feature of this dissertation is its focus on ideals of proof.

Our second observation is that mathematics as well as logic deal with an inherent interplay between formal and informal notions, for which the right trade-off needs to be found. For instance, "[t]hough formal rigor is crucial, it is not sufficient to shape proofs intelligibly or to discover them efficiently, even in pure logic" (Rathjen and Sieg, 2024). Hence, mathematics and logic switch between

[1] See for instance the news item https://www.theguardian.com/us-news/2023/mar/24/new-orleans-pythagoras-theorem-trigonometry-prove.

various levels of formality, and unavoidably deal with the formalization of informal notions. Still, how formalization works, and what are exactly the differences between informal and formal concepts, are not well-understood. A second main feature of this dissertation is its interest in formalization of notions surrounding mathematical proofs.

Bringing these two observations together, our aim is to define and compare several formalizations of ideals of proof. The ideals we consider will lack exactly the type of rigor that mathematical definitions are commonly thought to possess, yet they play a legitimate role in guiding mathematical practice. By studying their formalizations, we aim to better understand the nature of ideals of proof, as well as the nature of formalization (specifically on the level of formal proof systems). There has been relatively little attention for this topic in the literature. An example consists of Hilbert's '24$^{\text{th}}$ problem', which was left out of the famous list of 23 problems in mathematics that Hilbert presented at the International Congress of Mathematicians in Paris in 1900.

> "The 24th problem in my Paris lecture was to be: Criteria of simplicity, or proof of the greatest simplicity of certain proofs. Develop a theory of the method of proof in mathematics in general. Under a given set of conditions there can be but one simplest proof. Quite generally, if there are two proofs for a theorem, you must keep going until you have derived each from the other, or until it becomes quite evident what variant conditions (and aids) have been used in the two proofs. Given two routes, it is not right to take either of these two or to look for a third; it is necessary to investigate the area lying between the two routes." (English translation by Thiele (2003))

Some proposals have been given in the literature for criteria for simplicity (see e.g. (Hipolito and Kahle, 2019)), but the philosophical nature of the informal property of simplicity has shown that formalization is a complex process. We intend for our case studies to take concrete steps in the development of formalizations, specifically that of proof ideals.

## Main contributions

The main results of this dissertation concern three ideals of proof: that of purity (as opposed to impurity), explanation (concerning proofs that explain, versus proofs that do not), and semantic pollution (as opposed to syntactic purity).[2] These ideals will be properly introduced in Chapter 1. A first contribution of this dissertation

[2] A more appropriate description might mention instead of semantic pollution, its counterpart syntactic purity as the real 'ideal'. However, we remain relatively neutral on the value of these properties, and assume that either extreme of an 'ideal' of proof can become desirable in certain contexts and with certain purposes. Additionally, the emphasis in this dissertation will lie on providing measures of semantic pollution, instead of syntactic purity.

is then its relatively large-scale comparison of various models of purity and explanation (for informal proofs), relative to a case study of two particular informal proofs. Included in this comparison, and forming the second main contribution of this dissertation, is the introduction of a new model of purity to the literature, namely that of *ontological purity*. We define this model of purity not only on the level of informal proofs, but additionally for formalized (natural deduction) proofs as well. In doing this, we show that the combination of mathematical tools (such as Visser (1997)'s interpretation translation) with philosophical perspectives (like mathematical structuralism (Shapiro, 1997)) can lead to successful formalizations.

Third, we introduce an elaborate conceptual and formal understanding of semantic pollution, an ideal that is mentioned widely throughout proof-theoretic literature in connection to 'labeled' proof systems, but has not often been considered closely. This involves a case study of proof systems for modal logic, in particular various generalizations of usual sequent calculi — as well as accompanying Kripke models and the notions of equivalence that relate them. We take a range of proof-theoretic languages as objects of study, and we report a first systematic analysis of these languages under our measures of semantic pollution (where labeled languages emerge as relatively highly semantically polluted).

**Remark** (Use of the term 'purity'). The term 'purity' has two main uses. First, it occurs as an ideal of mathematical proof, and it will be used for specifying more specific variants such as topical purity, operational purity and ontological purity. Second, the term is used in 'syntactic purity' as the counterpart of semantic pollution. We will largely restrict to using the generic term 'purity' only for the general ideal of mathematical (informal) proofs. More specific terms such as 'ontological purity' and 'syntactic purity' will be used in these specific settings, possibly replaced by 'purity' only in cases where the intended interpretation is unmistakable from context.

**Remark** (Background knowledge). We assume that the reader has basic familiarity with the syntax, proof theory and semantics of the usual propositional, first-order and modal logics. Where relevant, specific definitions will be given, but we will often restrict to the level of detail that is relevant for our specific investigations. For instance, in discussing proof theory for modal logics in Chapters 5 and 6, we will not introduce proof systems by their full set of axioms and inference rules, but we focus instead on the grammar, the aspect that is most relevant to our subsequent formalization of semantic pollution. Of course, we always provide more specific references where appropriate. To start, broad and more rigorous formal treatments on natural deduction and sequent calculi can be found in (Negri and Von Plato, 2008; Buss, 1998), and a good coverage of the syntax and semantics for modal logic is (Blackburn et al., 2001).

## Structure of the thesis

The dissertation consists of seven chapters, and is intended to be read from beginning to end. However, several parts may be read independently, since we concern ourselves with different case studies.

First, Chapter 1 concerns a broad introduction to the case studies of the following chapters, and serves to define the wider context we can embed them in. There, we will introduce the distinction between informal and formal proofs, several aspects of formalizations, and the most well-known ideals for each type of proof.

Afterwards, Chapter 2 stands independently as a case study comparing models of purity and explanation for informal proofs. A preview of the notion of ontological purity can be found there, while it is presented in detail in Chapters 3 and 4, the latter which should be read in this order.

Furthermore, we create a formalization of semantic pollution (as opposed to syntactic purity) of formal proofs in Chapter 6, preceded by conceptual and technical preliminaries in Chapter 5 (that is thus intended to be read before Chapter 6). Zooming out, the case studies can be seen to concern mathematical proofs (Chapters 2, 3, 4) as well as logical proofs (Chapters 5, 6); and informal proofs (Chapters 2, 3, 4) as well as formal proofs (Chapters 3, 4, 5, 6).

Chapter 7 will take the previous chapters together in observations about the similarities and differences between our case studies, generalizations of our formalizations, and reflections on formalization of ideals generally. We end the dissertation with some concluding remarks.

## Sources of the material

Chapters 1 and 7 were written specifically for this dissertation. The other chapters are based on single- or co-authored papers, either published or submitted.

- Chapter 2 (in particular, Sections 2.1.1, 2.2.1, 2.3, 2.4, 2.5, and 2.6.3) is based on a collaboration with Francesca Poggiolesi:

  > R. Martinot and F. Poggiolesi. Purity and explanation: A systematic case study. *(Submitted)*, 2024

  The authors contributed equally to the mentioned sections. The other sections were added separately to the chapter, in order to better fit the story of the dissertation.

- Chapter 3 and 4 are published as one paper, as follows.

  > R. Martinot. Ontological purity for formal proofs. *The Review of Symbolic Logic*, 17(2):395–434, 2024a

- Chapters 5 (in particular, Sections 5.2, 5.3, and 5.4.1) and Chapter 6 are currently submitted as one paper:

  > R. Martinot. A formal characterization of semantic pollution of modal proof systems. *(Submitted)*, 2024b

- Chapter 5, Section 5.4.2 is based on parts of the following publication.

  > R. Martinot. Towards a formal analysis of semantic pollution of proof systems. *The Logica Yearbook 2022*, 33(1):79–98, 2022

# 1

# Ideals of proof (systems) and formalization

This chapter will provide an introduction into the various available ideals of informal (mathematical) proofs, and those of formal proofs (and proof systems). They range from ideals with a relatively technical focus, to ideals with a philosophical foundation, and they can often be traced back to historical origins. Our own interests in ideals of proof are motivated not only by a general interest in why mathematicians favor certain proofs over others, but also by ideals as case studies of *formalization*. In the chapters that follow, we will consider several formalizations (on different levels) of ideals of proof and proof systems. Here, we provide a general embedding of these case studies in the literature.

First, since both informal and formal proofs play a large role in this dissertation, we will consider their distinction in Section 1.1. We will then discuss the nature of formalization in Section 1.2, and formalization of ideals in Section 1.3. Subsequently, Sections 1.4 and 1.5 consider the presentation of ideals of mathematical (informal) proof, and ideals of formal proofs (and proof systems), respectively. We conclude in Section 1.6.

## 1.1 The relation between informal and formal proofs

The terms 'informal proof' and 'formal proof' may be used in several different ways. One interpretation of the distinction concerns their degree of satisfying a certain standard of rigor upheld by the mathematical community, where informal proofs are somehow 'faulty' and unreliable — this is not the interpretation we aim at here. Instead, the difference in formality is identified within the class of valid proofs. Here, informal proofs are mathematical proofs the way mathematicians write them down and use them in practice. As a consequence, they are commonly characterized as "linguistic entities" that "contain gaps of reasoning and appeals

to intuition" (Antonutti Marfori, 2010). One might even view them as "social conventions by which mathematicians convince one another of the truth of theorems" (as 'social proofs', (Buss, 1998, p.2)). Note that this means that the standards for what is an informal proof may fluctuate over time, and may fluctuate depending on the audience of the proof. Instead, *formal* proofs are formal in a precise sense, as they correspond to "a string of symbols which satisfy some precisely stated set of rules and which prove a theorem, which itself must also be expressed as a string of symbols" (Buss, 1998). The field of proof theory is dedicated to treating and studying formal proofs as mathematical objects — to a lesser extent, informal provability has also been the subject of formal treatment (see e.g. (Leitgeb, 2009)).

The notion of formal proof can be traced back to the emergence of *Hilbert's program* in 1921. With this foundational research program, Hilbert aimed to formalize classical mathematics entirely in axiomatic systems, as well as prove the consistency of these systems using only 'finitary' means. This was supposed to lead to a "philosophically satisfactory grounding of classical, infinitary mathematics (analysis and set theory)" (Zach, 2007). However, the infeasibility of Hilbert's program is typically regarded as having been shown by Gödel's Incompleteness Theorems. In particular, the Second Incompleteness Theorem shows that every consistent and effectively axiomatized theory that allows for the development of basic parts of arithmetic cannot prove its own consistency (see (Zach, 2007, 2023) for more discussion on the impact of this result on Hilbert's program).

Still, Hilbert's program led to numerous successful research areas, and proof theory itself may be considered a consequence of it. The axiomatic systems in which Hilbert aimed to formalize mathematics became known as *Hilbert-style proof systems*. In general, the notion of a proof system is known nowadays as containing, relative to a formal grammar, a collection of axiom schemes (whose substitution instances are axioms), and a collection of inference rules (that provide a way of deriving new formulas from previously derived ones). Hilbert-style proof systems are, in the spirit of Hilbert's aims, focused on finding the right axioms of a proof. Such a proof system comes with an abundance of axiom schemes, and a restricted number of inference rules. The standard Hilbert systems only contain the rule of Modus Ponens (MP).

$$\frac{A \qquad A \to B}{B}\ \mathsf{MP}$$

Two different styles of proof systems were introduced by, among others, Gentzen (1935) (see also (Gentzen, 1969)), who was primarily motivated by the aim of proving the consistency of arithmetic. The system of *natural deduction* resulted from the idea that a proof system should reflect more the way we reason in practice (it was independently also introduced by (Jaśkowski, 1934), who used graphical and linear styles of proof representation). Unlike Hilbert systems, natural deduction systems have in fact no axioms, but incorporate logical constants in terms of

inference rules. Additionally, natural deduction systems work with assumptions from which one can reason, that can be discarded once they are no longer necessary. Some of the natural deduction rules for classical propositional logic are the following, coming in the form of *introduction rules* (allowing a logical constant to appear in the conclusion) and *elimination rules* (allowing a logical constant to disappear in the conclusion).

$$\frac{A \quad B}{A \wedge B}\,\wedge\mathsf{I} \qquad \frac{A \wedge B}{A}\,\wedge\mathsf{E} \qquad \frac{A \wedge B}{B}\,\wedge\mathsf{E}$$

$$\frac{\begin{array}{c}[A]\\ \vdots \\ B\end{array}}{A \to B}\,\to\mathsf{I} \qquad \frac{A \to B \quad A}{B}\,\to\mathsf{E}$$

After believing that classical natural deduction could not provide suitable normalization results (a way of showing that the proofs in a proof system are not 'roundabout'), Gentzen developed a different style of proof systems known as the *sequent calculus*.

Sequent calculi keep track of the open assumptions in a proof by presenting them in a list in every proof line. The data structure occupying a proof line is no longer a formula, but a *sequent* $\Gamma \Rightarrow \Delta$, where (antecedent) $\Gamma$ and (succedent) $\Delta$ represent (originally) lists of logical formulas $A_1, ..., A_n$. Sequent calculi thus introduce the comma ',' as a new syntactic symbol, as well as the 'sequent arrow' $\Rightarrow$, which can (among others) be considered a formal representation of the derivability relation. The exact properties of the sequent ingredients may differ depending on their applications — for instance, there exist single-conclusion and multiple-conclusion sequent calculi, and one may vary the nature of $\Gamma$ and $\Delta$ (treating them as sets, multisets, sequences, lists or formulas). Logical constants are incorporated in a sequent calculus via 'left rules' (corresponding to natural deduction elimination rules) and 'right rules' (corresponding to natural deduction introduction rules). Some examples of inference rules from a multiple-conclusion sequent calculus are given below.

$$\frac{\Gamma, A, B \Rightarrow \Delta}{\Gamma, A \wedge B \Rightarrow \Delta}\,\wedge\mathsf{L} \qquad \frac{\Gamma \Rightarrow A, \Delta \quad \Gamma \Rightarrow B, \Delta}{\Gamma \Rightarrow A \wedge B, \Delta}\,\wedge\mathsf{R}$$

$$\frac{\Gamma \Rightarrow \Delta, A \quad \Gamma, B \Rightarrow \Delta}{\Gamma, A \to B \Rightarrow \Delta}\,\to\mathsf{L} \qquad \frac{\Gamma, A \Rightarrow B, \Delta}{\Gamma \Rightarrow A \to B, \Delta}\,\to\mathsf{R}$$

These types of proof systems comprise the three main proof-theoretic formulations of systems of logical rules, and they each provide a basis for generating 'formal proofs' (we refer to (Buss, 1998; Negri and Von Plato, 2008) for more details).

At this point, the question naturally arises how informal proofs and formal proofs relate to each other. This is an area where many questions currently still

exist, although several relations are generally agreed upon. For instance, as a consequence of their high rigor, formal proofs may serve as a *correctness check* for informal proofs (see e.g. (Avigad, 2021)). The difference in rigor, however, goes together with a difference in epistemological value of the two types of proof. It is generally recognized that formal proofs "are abstractions from mathematical practice that fail to capture many important aspects of that practice" (Dawson, 2006), and so unlike informal proofs, formal proofs do not generally convey a higher-level conceptual understanding of the reasoning used in a proof. Still, some are tempted to conjecture more specific relations between informal and formal proofs. The 'derivation-indicator view' (Azzouni, 2004) says that each informal proof 'indicates' an underlying formal derivation. How this is supposed to work, however, is unclear, and the idea is met with criticism in, among others, (Rav, 2007). Still, there are ongoing attempts to bridge the gap between informal and formal proofs, see for instance discussions in (Jojgov et al., 2004; Burgess, 2015; Weir, 2016), although a detailed understanding of how formal proofs more conceptually approximate informal proofs, remains largely unclear.

A better understanding of this approximation is desirable for various reasons, however. For instance, "[i]n order to make the wonderful tools from proof theory, model theory, recursion theory, set theory, modal logic, and so forth, applicable to the analysis of informal provability, we need bridge principles which relate informal provability to formal provability" Leitgeb (2009). Such 'bridge principles' would thus create a better understanding of informal provability itself, and it would allow justified transfer of results between the two types of proofs. Leitgeb notes that "[e]ven for giving partial answers to questions such as 'What do the Incompleteness Theorems tell us about mathematical provability?' we need a theory of informal provability".

Additionally, finding ways to preserve the informal intuition behind a formalized proof enables a more complete justification for the validity of a proof. As Hamami (2019) notes, formal deductions have a legitimizing role for human knowledge, and a routine translation of reasoning into formal deductions would contribute to the conceptual part of the justification. And more philosophically, once an area of mathematics has been formalized, the question remains: "[h]ow can we be sure that the formal system accurately reflects the original mathematical structures?" (Shapiro, 2006) (see also (Lakatos, 1978)). A better understanding of the relation between informal and formal proofs could help assess the suitability and naturalness of formalizations.

We also emphasize that investigating informal and formal provability has a tight connection to investigating the notion of formalization, generally. A better understanding of the relationship between informal and formal proofs contributes to characterizing the type of formalization that informal proofs are subject to. Vice versa, knowing more about formalization itself may contribute to bridging the gap between informal and formal provability. The following section will clarify more the type of formalization we will concern ourselves with in the next chapters.

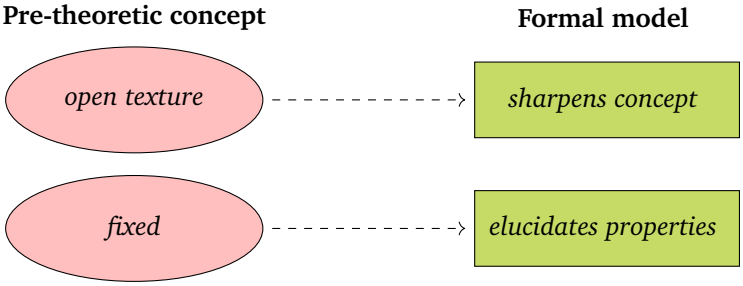



Figure 1.1: A visualization of two views on formalization.

## 1.2 Formalization

We base our general understanding of the notion of formalization on the ideas of Waismann (1968) (and analyses in (Shapiro, 2006; Incurvati, 2020)) and Hansson (2000). Let an informal concept that will be subject to formalization (into a formal model) be called the *pre-theoretic concept*. It is then recognized that we can distinguish between different views on the nature of the pre-theory — Incurvati (2020) describes three such views, two of which we take as a basis (see Figure 1.1). Generally, the first view says that pre-theoretic concepts are themselves imprecise, and formalizations serve to *sharpen* them, by clarifying properties that may not originally have been part of the concept. The second view claims that the pre-theoretic concept is itself in fact already sharp and fixed 'in all directions' — we just have a limited informal understanding of it. In this case, formalizations serve to *elucidate* the more detailed properties that were already part of the pre-theory, but that we did not perceive.[1]

Waismann (1968) is a main promoter of the first view. He argues that informal concepts possess *open texture*, which refers to the idea that "[t]he fact that in many cases there is no such thing as a conclusive verification is connected to the fact that most of our empirical concepts are not delimited in all possible directions" (Waismann, 1968). Although this is meant as an attack on phenomenalism (concerning the verification of sentences like *'there is a cat next door'*), Shapiro (2006) provides several mathematically flavored examples.

A well-known case is the successful formalization of the notion of computability. Church's Thesis proposes that all and only recursive functions are effectively computable — the intuitive concept of computability is thereby formalized into a rigorous one. This formalization involved fixing certain parameters surrounding

[1] Incurvati (2020)'s third view says that the pre-theoretic concept may be inconsistent, which is when a formalization serves to *replace* the concept by a new (consistent) one. This is a relevant view when the other two views assume that the pre-theory is always consistent. We will, however, assume that the possibility of inconsistency of the pre-theoretic concept is already present in the other two views (enabled by open texture, or our inaccurate perception of a fixed pre-theory).

the informal notion of computability, that are in principle up to interpretation. This includes "such matters as the attention span, lifetime [of a computist], and available materials" (Shapiro, 2006), and to what extent these should be idealized in order to end up with an interesting notion of computability. For instance, the amount of memory or materials available to a computist should not affect the possibility of a computation in order for it to be an interesting notion. According to Shapiro, by setting such parameters (based on mathematical efforts) the formalization of computability in fact *sharpened* its pre-theoretic version.

Another example concerns the concept of set. Shapiro (2006) notes that "[t]here is still some debate over the intuitive underpinning of the iterative conception", and Boolos (1989) argues there is not one single, informal notion of set underlying ZF. See also (Incurvati, 2020) for an elaborate analysis of different conceptions of set (neutrally taken as either 'sharpening', 'elucidating' or 'replacing' pre-theoretic concepts of set). Finally, take the example of the notion of a polyhedron. The teacher-student discussions described in (Lakatos, 2015) show that the pre-theoretic notion of a polyhedron is subject to interpretation and changing boundaries. Various definitions are proposed and rejected, illustrating the different ways one can interpret formalization of intuitive concepts.

Furthermore, a view common to Waismann (1968) and Hansson (2000), independent of the supposed 'sharpness' of the pre-theory, is that formalizations generally only sharpen or elucidate a concept on *some* aspects. That is, the pre-theoretic concept can never be 'completely' captured, thereby overlooking some aspects that perhaps were present in the pre-theory. Hansson (2000) calls attention to this by pointing out the 'simplifying' side to formalization (of rational behavior).

> "[F]ormal models of rational behaviour and rational belief [...] are both (1) idealizing — simplifying, i.e. they leave out many of the complexities of real life, and (2) idealizing — perfecting, i.e., they represent patterns that satisfy standards of rationality that are higher than what actual (doxastic) agents usually live up to." (Hansson, 2000)

The simplifying side to formalization has the purpose of obtaining at least *some* understanding of the concept under analysis. However, "[i]t may involve a distortion of the original or it can simply mean a leaving aside of some components in a complex in order to focus the better on the remaining ones" (McMullin, 1985).

We will thus take formalizations as multi-sided, not only sharpening or elucidating a pre-theoretic concept, but also possibly (over)simplifying it. It is useful to outline more specifically some of the possible virtues as well as dangers of formalization in philosophy that Hansson (2000) describes. Besides noting virtues that are similar to sharpening and elucidating aspects of the pretheory (e.g., isolating important aspects that underlie a concept, and shedding light on implicit assumptions made informally), Hansson notes that formalization also encourages "definitional and deductive economy" (such as by showing which concepts are in fact interdefinable or equivalent in the formal setting, and by minimizing the

principles of inference used). Additionally, it encourages a deeper understanding of concepts by supporting intricate theoretical frameworks that in the informal setting become too ambiguous to uphold; finally, formalization can lead to the discovery of previously unnoticed philosophical problems, by supporting a 'complete' formal description of an informal phenomenon.

On the other hand, Hansson draws attention to cases where formalization causes more confusion than clarity. We already mentioned that formalization can induce oversimplification. More specifically, formalization can cause false unification of informal concepts (into one formal notion); and it can propose technically pleasing but practically false conceptual primitives. On the other hand, formalization can also introduce unnecessary *complexity*. For instance, it can encourage introducing ad hoc constructions that have no natural informal counterpart; and it might lead one to focus on (philosophically irrelevant) problems that are in fact just technical artifacts of the formalization. Hansson finally mentions that formalization can involve (harmful) implicit ontological assumptions; and that rash (unmotivated) philosophical choices can be made by making formally convenient choices.

In short, to assess the use of formalizing, one should always be aware of which aspects of formalization contribute to which aspects of informal issues. We will now consider what views on formalization are (ir)relevant for the formalization of ideals of proof (systems) that will follow in the next chapters.

## 1.3 Formalization of ideals of proof (systems)

The kind of formalization that we will concern ourselves with focuses on ideals of proof (systems) as pre-theoretic concepts. The formalization of ideals of proof is an area that has not yet enjoyed many substantial results, but which relates for example to 'Hilbert's 24$^{\text{th}}$ problem' for formal proofs: the problem of finding criteria for simple proofs (Hipolito and Kahle, 2019). This involves formalization of particular *aspects* of these proofs, namely specific conceptual qualities that they possess. In particular, our formalizations will concern the following cases, taking place on different levels of analysis.

- Formalizing pre-theoretic intuitions about *an ideal of informal proofs*, into a model of the ideal for informal proofs (Chapter 2). The relevant ideals in this chapter will be *purity* and *explanation*.
- Formalizing pre-theoretic intuitions about an ideal of informal proofs, into a model of the ideal for formal proofs (Chapters 3 and 4). More specifically, this formalization will concern *first-order natural deduction proofs*, and the relevant ideal will be *purity*.
- Formalizing pre-theoretic intuitions about an ideal of formal proofs, into a model of the ideal for formal proofs (Chapters 5 and 6). More specifically,

> this formalization will concern proofs of (generalizations of) *propositional modal sequent calculi*, and the relevant ideal will be *syntactic purity* (as opposed to *semantic pollution*).

Thus, models of ideals of proof will look quite different, depending on the type of proof that the ideal and formalization concern itself with. Generally, on the level of informal as well as formal proofs, a model should specify under what conditions a proof can be said to satisfy the ideal, as well as motivate how these conditions relate to the pre-theoretic understanding of the ideal. On the level of formal proofs, we expect there to be a relation of the ideal to rigid counterparts of formal proofs (such as syntax or inference rules). Furthermore, given the described positions, virtues and dangers of formalization, we will assume that the type of formalizations we will use for ideals of proof have at least the following properties, that we elaborate on below.

1. We will adhere to *pluralism* for formalizations.
2. We remain *neutral* on whether a formalization 'sharpens' a pre-theory with open texture, or whether it 'elucidates' aspects of a fixed pre-theory.
3. We will assume formalizations to have both virtuous as well as harmful sides.

Key to our approach to formalizations is that we consider there to be multiple 'acceptable' formalizations of ideals of proof. Depending on the goal one has, one can pick a particular formalization that has the right properties — giving rise to a pluralist attitude towards formalizations. Note that this seems especially compatible with the *open texture* view on pre-theoretical ideals: the pre-theory will contain undetermined aspects, and formalizations can 'sharpen' them in different ways. This is compatible with the following idea:

> "[I]t should be clear that there is no unique "true" formal analysis of non-philosophical or informal philosophical concepts. Different formalizations may capture different properties of the concepts." (Hansson, 2000)

Even on the view of a 'fixed pre-theory', however, pluralism with respect to formalizations is acceptable. Different formalizations may still 'elucidate' different aspects of the (fixed) pre-theory, and ignore others.

A basic consequence of pluralism of formalizations is that different formalizations (of the same proof ideal) can lead to different judgements on whether a proof possesses the ideal, or not. We might expect this to only be compatible with the 'open texture view' of pre-theories. It seems natural that different choices of sharpening the pre-theory can lead to different outcomes for the same proof. On the 'fixed pre-theory' view, we might expect at first sight that different formalizations should provide compatible judgements, as they describe the same

(fixed) pre-theory. However, recall that two formalizations of the fixed pre-theory may still focus on elucidating *some* (different) aspects, and ignore the elucidation of others. In case a proof possesses an aspect that only one formalization elucidates, two formalizations might still provide different judgements. Thus, we have reached our second assumption, which is that we remain neutral with respect to the nature of the pre-theory.[2]

Third, pluralism on formalizations is naturally compatible with a nuanced view on the value of formalizations. By sharpening or elucidating 'some' and not all properties of an ideal, a formalization inevitably simplifies or ignores others. We may thus assume that any formalization is biased to a certain extent to the properties that it focuses on, thereby providing a 'distorted' view of the pre-theory. Generally, then, we expect that 'Lakatosian monsters' may appear, that satisfy the formalizations, yet seem counter-intuitive. It should be stressed that we do not consider this to be a reason to reject the formalism. In the spirit of Hansson (2000), "even if such a counter-argument convincingly discloses a deficiency in the model, this is not necessarily sufficient reason to give up the model. If the counter-argument be neutralized without substantial losses of simplicity, then an appropriate response may be to continue using the model, bearing in mind its weaknesses". We can consider such examples to belong to the ideal of proof relative to the aspects that have been formalized — and we might consider the formalization to focus on these aspects of a proof ideal, instead of on the entire ideal itself.

Finally, despite the fact that we advocate a plurality of formalizations, we recognize that different pre-theoretic concepts will fit a monist or pluralist approach to formalization best. In the case of computability, Shapiro (2006) considered the formalization to essentially *remove* the open texture of the concept, so that mathematicians could settle on a 'right' concept. If there are natural choices for sharpenings such that the mathematical community largely agrees on them, then pluralism may turn into monism, and we do not claim that this cannot happen with certain ideals of proof. Alternatively, if pluralism of the formalization of some ideal causes long-term disagreements in the mathematical community, differing formalizations might instead be taken to be formalize *different* ideals, splitting up a pre-theoretic ideal into multiple ones. Such choices, however, all depend on the development of mathematical practice, and for now we may simply adopt pluralism towards formalizations.

[2]Despite this neutrality, a more biased view might be more appropriate depending on the particular proof ideal. An open-texture pre-theory seems more likely relative to a high-level, philosophical ideal — whereas our intuitive understanding of some ideals of formal proofs seem from the beginning already more rigid.

### 1.3.1 The value of considering such formalizations of ideals

The reader may still be wondering why our perspective on formalizations of ideals is an interesting one to take, and what this implies about ideals of proof themselves. We will thus briefly elaborate on the value of ideals of proof (relative to the properties of formalization presented above), as well as on the value of the formalizations themselves.

Concerning the first, we recognize that originally, ideals of (informal) proof were intended (by scholars that discussed them, such as Aristotle and Bolzano on 'purity' or explanatory value of proofs) to select the 'highest quality' proof, a seemingly monist endeavor. The ideal of proof then favors a proof possessing the ideal, over a proof not possessing the ideal. Arguably, a *formalization* of an ideal is intended to clarify more precisely what the ideal amounts to, and should help us detect whether or not a proof possesses it. However, adhering to pluralism of formalizations now gives us multiple ways of judging a proof on the ideal, and these ways may be inconsistent. This seems to change the role of ideals, as their formalizations cannot generally select one, highest-quality proof.

However, as in line with the previous section, we argue that ideals of proof still select high-quality proofs. They may just be considered to do this relative to particular *aspects* of the ideal. We may choose a particular formalization as guiding the choice of 'best proof' relative to our goal that favors these aspects. A goal of a formalization may be something technical and specific, or something more general and philosophical: it may even concern aspects from *different* ideals of proof (instead of finding an explanatory proof, one might want to find a *fruitful* explanatory proof, for which a formalization of explanation emphasizing connections to other results may provide the best guidance). Ideals of proof thus become a collection of favorable aspects, that show us more precisely the different ways in which mathematics can be informative and valuable. [3]

Now, why are we interested in studying formalizations of ideals of proof themselves? First of all, naturally, formalizations provide insights into (aspects of) ideals of proof. Tappenden (2012) notes that analyses of ideals of proof can "clarify what it is that makes mathematics informative and revealing". Formalization may thus help us see why mathematicians prefer certain valid proofs over others, and help explain mathematical practice. Related is the observation that ideals of proof, as properties of desirable proof, can be seen as guiding the search for proofs generally. This raises a question: "*[h]ow* can such pragmatic and apparently subjective advantages possibly serve as reliable guides to what is *true*?" (Tappenden, 2012). It remains unclear why philosophical properties of proof, if at all, are related to mathematical success. "The methodologist has the task of clarifying the nature

[3] And if, amidst all formalizations of a certain ideal, one insists on knowing which one is the 'best' or 'true' formalization capturing the ideal — one might instead be willing to take each formalization as 'creating' a new (variant of the) proof ideal. The rivalry between formalizations then once again corresponds to a rivalry between proof ideals, which was already there.

and role of these theoretical virtues, or demonstrating that they do not guide theory choice in a significant way after all, and that appearances to the contrary are misleading" (Tappenden, 2012).

Furthermore, formalizations of ideals can be considered case studies of the notion of formalization, generally. Section 1.1 already illustrated that the process of formalization in mathematics is not well-understood: in particular, this concerns the relation of informal proofs to formal proofs, but also the more general precisification of informal properties of (informal or formal) proofs. Studying and devising specific formalizations should thus contribute more generally to understanding what is a (successful) formalization.

Finally, formalizations of ideals of proof can tell us several things about mathematical practice. For one, a successful formalization tells us that there are relatively consistent frameworks underlying our (informal) mathematical intuitions; while it also shows us which parts of our intuitions may not be consistent. This knowledge can help shape the philosophical debates surrounding ideals. Formalizations can pinpoint where disagreements originate, and can help us show in which situations we we should (not) consider certain aspects of ideals valuable. Formalization then prevents miscommunication when ideals of proof are understood in different ways by different mathematicians or philosophers.

The rest of this chapter will introduce the main ideals of proof (systems) that are mentioned in the literature. In the chapters that follow, we will dive into various formalizations of specific ideals of proof. After that, we will reflect on what is gained as well as lost by the formalizations in Chapter 7.[4]

## 1.4 Ideals of informal (mathematical) proof

In order to obtain a better understanding of the nature of ideals of informal, mathematical proofs, let us name the most common ones. In the philosophy of mathematics, ideals of informal proofs are also called 'virtues' of proof, and a proof may possess various types of them. Lange (2016b) mentions the following list: *accessibility* to a given audience, *beauty* (see e.g. (Novaes, 2019)), *brevity*, *depth*, *elegance*, *explanatory power* (with a wide literature, see also Chapter 2), *fruitfulness*, *generalizability*, *purity* (see Chapters 2, 3, 4), and *visualizability*. Even others are "standards of *rigor*, *certainty*, *apriority*"[5], and *simplicity* of proof (see (Daw-

[4]Note that, even though we introduce specific formalizations of proof ideals, this does not contradict our pluralist conception of formalization. We consider our own formalizations to focus on particular aspects that we find interesting relative to the current mathematical and philosophical debate. However, other formalizations may still be more appropriate relative to other goals or other aspects of the pre-theory.

[5]See https://mdetlefsen.nd.edu/research/ideals-of-proof-ip/.

son, 2006; Hipolito and Kahle, 2019)).[6] Often, mathematicians and philosophers only possess a rudimentary, intuitive understanding of such an ideal. For instance, we might say that elegance of a proof holds when "[a] proof (and so on) can be praised for attaining in a more efficient and "clean" way what previously had been done indirectly and clumsily" (Tappenden, 2012), but how exactly to be more precise is unclear, and seems best illustrated by concrete examples. Common to all ideals of proof is that the judgements of mathematicians are key in establishing whether the ideal holds. Although unavoidable, this can be problematic when narrowing down the concepts behind ideals of proof: it can already be unclear "whether mathematicians are appealing to the same virtue each time they invoke 'beauty' (for instance)" (Lange, 2016b). It depends on the particular ideal of proof whether other, external factors are also at play in determining whether a proof possesses it. For fruitfulness of mathematical proof, for instance, it is "not just a question of the feelings or subjective reactions of mathematicians, since there is a genuine fact of the matter at stake: will this proposal actually lead to the discoveries we are looking for?" (Tappenden, 2012).

This relates to a main distinction that can be made between the kinds of values that ideals of proof provide to mathematicians. First, ideals can be valuable mainly in an 'extrinsic' sense. Here, ideals ensure that a proof is useful in order to achieve other goals. For instance, accessibility to a given audience, brevity, and fruitfulness can be seen to possess this type of value: they help to communicate the proof to an audience, or to stimulate the discovery of new theorems or proofs. That is, ideals of proof are here *strategic* or *pragmatic* in the first place: such an ideal is "not so much prized for the knowledge it itself *constitutes*, as for the knowledge it in some broadly pragmatist sense provides for" (Detlefsen, 2008). But ideals of proof may also merely be 'intrinsically' valuable, so that "mathematics may seek them not as a means to some end, but as ends in themselves" (Lange, 2016b). Lange notes as examples here "explanatory power and purity, perhaps". Here, we may see ideals of proof as achieving higher epistemic goals: they are assumed to lead to a 'higher quality of knowledge' of a proof. On the other hand, we may imagine that some ideals of proof only possess 'aesthetic value' to mathematicians, without a clear relation even to epistemic values.

Hence, ideals of proof play a role in mathematical practice, guiding the search for 'good' proofs. Among their variety, however, there is not generally an order of importance to be recognized. I.e., "there need not be any sense in which a proof that exhibits one of them is 'better, all things considered' than a proof that lacks any of these virtues" (Lange, 2016b). Similarly, a proof of a theorem possessing one virtue is a valuable find, even if it lacks other virtues. The relation between ideals of proof can theoretically be many things: ideals of proof may be dependent on each other, (partly) overlap or be reducible to one another, but they may also

[6] As these terms are of a general nature, it would be interesting to see whether (in line with anti-exceptionalist views of logic) they may also be applicable to the way that we choose scientific theories, in general.

be entirely independent, or "they may stand in some more complicated relation" (Lange, 2016b). We will attempt to explore one such relation between two proof ideals in Chapter 2.

Finally, ideals of proof indirectly tell us something about 'sameness' of proofs. Namely, when a proof is considered to possess a certain ideal of proof, it is also considered to be sufficiently *conceptually distinct* from a proof that does not possess this ideal. Where the boundary lies between two proofs, i.e., when two proofs can be called *distinct*, is a non-obvious question (see, for instance, (Straßurger, 2007)). The case studies that we address in this dissertation can be seen as contributing to defining some boundaries (in certain contexts) — although we do not address the general problem of the identity of proofs.

Having said all this, the dynamics of mathematical practice may change ideals over time, and so the aspects by which a certain ideal of proof is known at one point in time, can be significantly different at another. For instance, purity was already by Aristotle considered a "quality of highest or best proof" (Detlefsen, 2008), but Detlefsen describes how the notion of purity received different conceptions over time. This should not be considered something problematic: in fact, it seems only appropriate with respect to a pluralist conception to formalizations of ideals.

## 1.5 Ideals of proof systems

Intuitively, like ideals of informal proofs, ideals of formal proofs are independent of provability. That is, given multiple ways of (formally) proving the same theorem, an ideal will for some reason favor one formal proof over the other. Unlike many ideals of informal proofs, however, ideals of formal proofs do not as such have to do with delivering the highest quality of 'knowledge' of the mathematical content of a proof — simply because a formal proof does not primarily serve to provide conceptual understanding of a proof. However, ideals of formal proof may serve other types of epistemic goals; and they often also serve pragmatic goals within mathematics. In general, a formal proof satisfying an ideal of proof will display a 'higher quality' of derivability than another, relative to these goals.

There are various levels at which we can present such ideals: we can phrase the ideal at the level of the *formal proof* itself, but we can also present it at the level of a *proof system*, or even more generally, at the level of a '*proof-theoretic formalism*'. Specific to sequent caluli, although we may use this term as applying to any type of proof system, a proof-theoretic formalism is given by "the standard notation it uses, the data structures employed in sequents, the types of inference rules that normally appear, and the types of properties ordinarily shared by the proof calculi thereof (which serve as instances of a formalism)" (Lyon et al., 2023). Ideals phrased at one level, can be seen to affect the other levels: an ideal of a proof-theoretic formalism can also be phrased from the perspective of a formal proof,

requiring that it should 'be part of a proof system that instantiates a proof-theoretic formalism with that ideal'; and the other way around. To present ideals of formal proofs, we will use the level that the ideal is most naturally phrased at.

In the literature, various lists of (slightly differing) desiderata of proof systems are provided (see e.g. (Avron, 1996; Wansing, 1994; Poggiolesi, 2010; Lyon et al., 2023)). We will distinguish between ideals of proof systems for the more *mathematically-minded* proof theorist, and ideals of proof systems for the more *philosophically-minded* proof theorist (although we recognize one may be a proof theorist with both interests). The mathematically-minded proof theorist is generally interested in ideals that help "generate large classes of proof calculi for logics on demand without requiring substantial work on the side of the logician" (Lyon et al., 2023). Additionally, ideals that help establish technical results about proof calculi fall under this category. On the other hand, the philosophically-minded proof theorist cares more about the 'rightness' of formal properties relative to a philosophical goal.

With this distinction in mind, it should be noted that compared to ideals of informal proofs, for formal proofs there is some more ambiguity regarding when something is an ideal of proof, as opposed to simply a 'mathematically desirable result' — as well as when something is a pre-theoretic ideal of proof, as opposed to a *formalization* of an ideal. The first ambiguity has to do with the fact that formal proofs and proof systems are themselves mathematical objects of study in proof theory. A technical desirable property of formal proofs may then simply be an open problem in proof theory, instead of a 'higher-level' quality of formal derivability. The second ambiguity has to do with the fact that the intuitive ideals of formal proofs are often already a bit more rigid, and that the formalization of such intuitions may sometimes proceed in multiple steps, the ideal becoming more formal or specific with each step, clouding the boundary between pre-theory and formalizations. We do not aim to provide an absolute boundary between these concepts for ideals of formal proofs — we will present several of them as presented by the literature, and simply advocate caution with respect to the terminology 'ideal of formal proof'.

First, then, consider some mathematically-oriented ideals, which are presented at the level of a proof-theoretic formalism as in (Lyon et al., 2023) (based on (Wansing, 1994; Avron, 1996)). Satisfying these ideals generally increases the ease with which proof systems can be generated and interpreted.

1. *Generality.* The proof-theoretic formalism that the proof system belongs to is not biased to particular logics: instead, it can deal with a large class of logics with proof systems (that share desirable properties).

2. *Uniformity.* The proof-theoretic formalism that the proof system belongs to, is stable the way it is: it does not need to be enriched in order to obtain proof systems for particular logics.

3. *Modularity.* A proof system for a logic in some class, can be transformed into a proof system for another "with properties preserved, by adding/deleting rules or modifying the functionality of rules" (Lyon et al., 2023).

4. *Constructibility.* The proof system for a logic in some class, is constructible according to an established method.

5. *Syntactic parsimony.*[7] The data structures that occur in the proof system are only as complex as they need to be, relative to the logic or (technical) purpose of the proof system.

The above ideals are presented on a relatively high conceptual level. Some well-known ideals at the level of proof systems for instance look as follows (again see (Lyon et al., 2023)), specifying desirable properties of concrete ingredients (or the formal proofs) of a calculus.

1. *Analyticity*. A formal proof is analytic if the premises of each rule only contain subformulas of the conclusion. This is a widely recognized value of formal proofs, relating to results like interpolation, the disjunction property, decidability, and so on. Analyticity in sequent calculi follows from establishing the possibility of cut-elimination, which says that each proof that incorporates uses of the cut rule, can be transformed into proof of the same theorem that is cut-free.

2. *Termination.* A proof system is terminating if the premises of each rule are less complex than the conclusion. Combined with analyticity, this property ensures (in sequent calculi) termination of proof search: when following rules bottom-up, every branch of a derivation leads to either an axiom or an unprovable sequent.

3. *Invertibility.* An inference rule of a formal proof is invertible if derivability of the conclusion of a rule implies derivability of its premises. This enables efficient proof search, as it ensures that it does not matter in which order rules are applied bottom-up during proof search.

4. *Countermodel generation.* Proof search in sequent calculi may provide the result that a conclusion is not provable, where the search ends with a derivation containing an unprovable sequent as a leaf. A calculus then generates countermodels if this sequent can be used to define a countermodel of the conclusion.

In the above cases, the ideals of formal proofs are so precise already, that their pre-theory and formalization almost coincide. We might also conceive of these properties instead as formalizations of a wider, higher-level ideal of formal proof

[7] We split Lyon et al. (2023)'s criterion of *syntactic parsimony* into a version of the criterion for mathematical reasons, and a version of the criterion for philosophical reasons.

such as *efficiency* or *applicability*, and that by force of familiarity they have started to be used as ideals in their own right. We will not elaborate further on them here.

For more philosophically-oriented ideals of formal proofs, we can more easily recognize a gap between the pre-theory and its formalization. Some examples of philosophical ideals are the following, ranging from desires for clear connections to informal reasoning, to requirements on the relationship of proof systems to meaning.

1. *Conceptual simplicity.*[8] The data structures that are used by the proof system are only as complex as required by the logic or (philosophical) purpose of the proof system. For instance, they are simple enough to sufficiently support a conceptual understanding of the (ingredients of the) data structure. This relates to the next ideal, but is here independent of inferential practice.

2. *Closeness to inferential practice*. This is a relevant requirement for a proof system if one is interested in understanding the relation between informal and formal reasoning better. It is related to Steinberger (2011)'s *Principle of Answerability* of proof systems, which says that "only such deductive systems are permissible as can be seen to be suitably connected to our ordinary inferential practices". Steinberger argues, for instance, that multiple-conclusion proof systems do not satisfy this principle.

   This ideal can manifest itself on different levels of analysis. We can distinguish between proof systems that aim to formalize entire ways of informal reasoning, and those that focus only on properties of informal reasoning. For example, some proof systems are intended to formalize deontic reasoning (e.g. (Torre and Villata, 2014)), or to focus on avoiding paradoxes such as logical omniscience. Then there are philosophical properties of mathematical reasoning such as our ideals of proof discussed in Section 1.4, and one might focus on formalizing just those aspects. On a more fundamental level, the type of proof system to begin with is thought to influence suitability of formalization. There, natural deduction is often mentioned as "[intending] to capture the way [we] actually reason" (Bimbó, 2014) (originally already noted by Gentzen), more than systems such as sequent or Hilbert calculi.

3. *Normativity.* Similar to desiring normativity of logic, we can seek normativity of a proof system. Here, we aim to trust the proof system to be a reasoning system that we can (and should) look to for improving our arguments (see e.g. (Tosatto et al., 2012)). Thus, instead of closeness to inferential practice, a proof system may instead be taken to display closeness to how we *ought* to reason.

[8] This can be seen as the philosophical variant of syntactic parsimony, which we dubbed 'conceptual simplicity' for clearer pre-theoretic nature.

4. *Providing meaning to logical constants. Inferentialism* is a broad philosophical framework (with the slogan 'meaning-as-use') according to which inferences establish the meaning of expressions (as opposed to 'denotationalism' as commonly used in model theory). *Proof-theoretic semantics* may be considered the more concrete efforts of creating inferentialist proof systems, that can provide the meaning of logical constants. A well-known worry to inferentialists is that we can devise rules for a connective *'tonk'*, that allow us to prove anything, and that render the connective meaningless. The common conclusion is that there should be constraints on inference rules to make them suitable for inferentialism.

5. *Categoricity*. A proof system is categorical if one can uniquely determine the truth conditions of logical constants from their proof-theoretic inference rules. This ideal can be considered a philosophical one in the sense that we want a proof system to properly connect to the semantics, and rule out 'unwanted' valuations. Carnap (1943) showed that basic proof systems already fail to satisfy this notion; however, there are various known methods for creating categorical proof systems (see e.g. (Smiley, 1996; Hjortland, 2014; Bonnay and Westerståhl, 2016, 2021)).

6. *Syntactic purity.* A proof system should not make any use of any (model-theoretic) 'semantic elements', for example, possible worlds or truth values (Avron, 1996; Poggiolesi, 2010). We will leave any more detailed interpretations of this ideal to the discussions in Chapter 5 and 6.

The less rigid pre-theories of many of these ideals leaves it unclear what formalizations could correspond to them. Proof-theoretic semantics is an interesting area of research where more and more formal properties of 'meaning-conferring' rules are identified. We name some of them below, which can thus be seen as further sharpening the fourth point in the above list. The first three properties ensure certain conditions for meaning-assignment of logical rules; together, they ensure that "every connective that is explicitly definable in [a given logic] L also has separated, symmetric, and explicit introduction rules" (Wansing, 1994).

1. *Explicitness*. The inference rules for a logical connective $C$ are *weakly explicit* if they exhibit $C$ in their conclusion sequents only. They are *explicit* if they are both weakly explicit, and the (relevant) rules exhibit only one occurrence of $C$ on the right and on the left.

2. *Separation*. Logical rules should only exhibit the constant that they introduce.

3. *Symmetry*. The rules for a connective $C$ are *weakly symmetric* if every rule either belongs to a set of rules that introduce $C$ into premises, or to a set of rules that introduce $C$ into conclusions. The rules are *symmetric* if they are both weakly symmetric and both types of rules are non-empty.

4. *Uniqueness.* A connective is *uniquely characterized* in a system that unifies two languages with the same connective, that denote it by a different syntactic symbol ($C$ and $C^*$) — iff for every formula in that unified language, $A(C)$ is provable in the system iff $A(C^*)$ is provable in the system.

5. *Harmony*. Harmony is a broad term for a requirement on the introduction and elimination rules being 'in balance', in order to somehow exclude 'pathological' connectives like Prior's *tonk* from counting as meaning-conferring. There is no unique accepted formal criterion for harmony, but various ones exist in the literature, and discussions on their relationship and suitability are ongoing (see, for instance, (Francez and Dyckhoff, 2012; Schroeder-Heister, 2014)).

In Chapter 5 we will see the way in which proof systems for modal logic are subject to enrichment of the data structure of sequents in order to achieve some of these desiderata.

## 1.6 Conclusion

In this chapter, we have provided the general embedding of the work following in the next chapters. We have seen that the relation between informal proofs as present in mathematical practice, and formal proofs as syntactic derivations in proof systems, is wanting for more clarity. Additionally, the notion of formalization is one that should be taken in a nuanced way, with regard for its benefits as well as dangers. We discussed a range of philosophical ideals of mathematical proof, as well as a variety of more technical ideals as well as philosophical ideals of formal proofs. The next chapters aim to put into practice the notion of formalization in various case studies of ideals of proof.

# 2

# Two ideals of proof

## *Purity and explanation: models and interactions*

We begin our investigation of ideals of proof at the level of *informal* mathematical proofs. Ideals or 'virtues' of informal proofs encompass a variety of properties, among them mathematical depth, beauty, simplicity, fruitfulness, elegance, and so on. We here select and zoom in on two such properties, namely purity and explanation. Both ideals of proof have historical roots (in works of, for instance, Aristotle and Bolzano), and there appears to be a connection in their origin. Still, the literature on purity of proof is growing and is especially wide in the case of explanation (see for instance (Lange, 2017)), and more recent works mainly highlight the differences between purity and explanation. This shows that the nature of ideals of proof changes over time, and that their value is closely tied to the dynamics of mathematical practice. In this chapter, we are interested in two things: first, what are the different ways in which purity and explanation of informal proof can be made precise? We will discuss various models of the two notions, including a preview of the notion of 'ontological purity', described in (Martinot, 2024a), which we discuss elaborately in Chapters 3 and 4. Second, we are interested in how the two ideals of purity and explanation interact, and we explore one way of approaching this question by taking several presented models of purity and explanation as a starting point.

First, Sections 2.1 and 2.2 will provide an intuitive introduction to the ideals of purity and explanation, as well as several examples of proofs possessing these values. Among them are the proofs of two theorems that will serve as case studies in the next part of the chapter, where we begin the comparison between purity and explanation. Section 2.3 provides some preliminary remarks on how our comparison will be carried out, followed by the actual comparison in Section 2.4 and 2.5.

We provide some more reflections in Section 2.6. This chapter corresponds to an adapted version of the submitted work of (Martinot and Poggiolesi, 2024).

## 2.1 Purity of proof

Purity has a long history as an ideal of proof for mathematicians, tracing back to early writings of Aristotle and Archimedes (Detlefsen, 2008). The notion concerns itself with certain restrictions in the way we are allowed to prove a theorem, and has several interpretations in the literature. Generally, we take a pure proof to only draw upon notions that belong to the content of the theorem. Here, 'content' should be taken as synonymous to the term 'topic' or 'subject matter'. Impure proofs distinguish themselves from pure proofs by making use of concepts that are somehow extraneous to what a theorem is about. To provide an intuitive sense of what purity involves, we mention two common types of (im)pure proofs found in the literature.

Dawson (2006) mentions that the search for purity can be recognized in attempts to exclude topological elements from proofs of the Fundamental Theorem of Algebra. We recognize this type of purity more generally, as well, for example in the idea that "[i]n many cases our understanding is not satisfied when, in a proof of a proposition of arithmetic, we appeal to geometry, or in proving a *geometrical truth* we draw on *function theory*" (a quote from Hilbert cited in (Hallett, 2007)). This illustrates a common view that a theorem belongs to a particular discipline of mathematics, and that the particular discipline used to prove the theorem in affects purity results. Similarly, a contrast is often made between arithmetical proofs of the Infinitude of Primes (IP), and a topological proof of the theorem by Furstenberg (1955) — we discuss this example soon in Section 2.1.1.

A different shade of purity statements suggests that impurity can occur when the theorem and proof still share a general conception of a mathematical sort (such as 'number' or 'set', where we here take a discipline of mathematics to concern itself with a specific sort). For instance, although all numerical domains can be thought to concern numbers, Dawson (2006) implies that impurity may occur when allowing imaginary values to occur in a real-valued power series. Additionally, consider Planar Desargues's Theorem, which assumes that two triangles $ABC$ and $A'B'C'$ lie in the same plane, and that they "are so arranged that the lines $AA'$, $BB'$, $CC'$ meet at a point. Desargues's Theorem then says that the three points of intersection generated by the three pairs of straight lines $AB$ and $A'B'$, $BC$ and $B'C'$, $AC$ and $A'C'$ themselves lie on a straight line" (Hallett, 2007) (see also Figure 2.1). A common proof of this theorem uses a point outside the plane, and draws upon all spatial axioms of Hilbert's incidence and order axioms (Groups I and II). Although the theorem and the proof share their focus on geometrical primitives such as points and lines, the theorem appears to only concern a two-dimensional version of these notions. Therefore, the point outside the plane

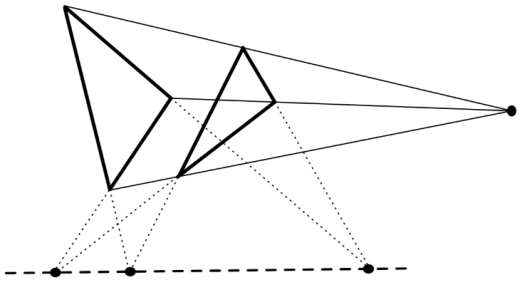


Figure 2.1: Visualization of Desargues's Theorem, adapted from (Arana and Mancosu, 2012).

can be seen as a source of impurity (see also (Arana and Mancosu, 2012)). A proof from just Hilbert's linear and planar axioms (Group I 1-2 and Group II 1-5), however, does not exist.

(Im)purity of proof has been considered to provide various values for mathematicians over time. Purity was originally taken to indicate reliability of proof. For one, a pure proof may be more appropriate to the theorem than an impure proof, as the latter "strays at points from its proper topic and is not in all its parts relevant" (Detlefsen, 2008). Additionally, pure proofs might match more the generality of the theorem they prove, an argument put forward originally by Bolzano (1817). This relates to the idea that different domains of mathematics should not be mixed, because there is a "natural order of priority among truths in mathematics and logic", a belief for instance supported by Frege (Burge et al., 2005). In this traditional conception, a pure proof does not rely on a subject matter that violates the right 'hierarchy' of truths.

While this idea has largely disappeared nowadays, remaining values of purity focus more on epistemic benefits. For example, pure proofs allow us to "become familiar with the specific details of the subject of the theorem" (Lehet, 2021). In other words, we acquire a 'deeper' understanding of the truth of a theorem based only on local, direct properties of the relevant objects. Lehet observes, however, that contemporary mathematics has lost much of its interest for pure proofs. Instead, the traditional conception of impurity fits today's mathematical endeavours better. Although impure proofs might divert the attention to notions outside the topic of the theorem, they are valued for their ability to unify and generalize mathematical results. For example, by proving analytic theorems algebraically, the distinct mathematical domains of analysis and algebra are unified by showing that they have intrinsic conceptual relations. Further, Lehet recognizes that explanatoriness can accompany impurity, for example when multiple objects from different mathematical disciplines are represented by the same structure in category theory. The relation between impurity and other epistemic values such as simplicity and explanatoriness is also explored by Arana (2017); Iemhoff (2017);

Lange (2019).

The main example that we will focus on in our case study, is the following.

### 2.1.1 The Infinitude of Primes

Consider the following theorem.

**2.1.1 Theorem.** ***Infinitude of Primes (***IP***).*** *For each natural number $n$, there exists a prime $p$ such that $p > n$.*

We present two proofs of this theorem, that present examples of purity and of impurity. Consider first the pure proof of the theorem of the Infinitude of Primes, originally presented by Euclid.

*Pure proof.* The proof proceeds by induction on $n$. For the base case $n = 1$, it is easy to take the prime $p = 2$. For the case $n > 1$, let $p_1, ..., p_m$ be all the primes less than or equal to $n$. Then, let $Q = (p_1 \cdot ... \cdot p_m) + 1$. If $Q$ is prime, then it is the prime we were looking for and the proof ends. If $Q$ is not prime, then, using the Fundamental Theorem of Arithmetic, we know that there is a prime $b$ that divides $Q$. We see that it must be the case that $b \neq p_i$ for any $i \leq m$, because if $b = p_i$ for some $i \leq m$, it would hold that $b$ divides $Q$ and also that $b$ divides $Q + 1$. Hence, $b$ divides $1$, a contradiction; so $b$ must in fact be a new prime. Also, $b \nleq n$ by the assumption that before we listed all primes lower than or equal to $n$, so $b > n$, and $b$ is the prime that we were looking for. □

Without already going into the details of specific models of purity, intuitively, pure proofs only use concepts that inherently belong to (sometimes explicated as the concepts that are 'mentioned by') the theorem being proved, and avoid the introduction of 'foreign' elements. As analyzed by Arana and Detlefsen (2011), Euclid's proof can naturally be carried out in Peano Arithmetic, and "it is reasonable to think of these axioms together with the [...] definitions of primality and divisibility as at least approximating the topic of IP" (see also mentionings in (Thorstad, 2012; Pillay, 2021; Kahle and Pulcini, 2017)). That is, the claim is that IP naturally concerns a collection of arithmetical concepts surrounding the natural numbers and primality. Euclid's proof, using only bounded sequences of primes, some instances of multiplication, addition and division, and an order relation, does not introduce any concept that seems extrinsic to this arithmetical context. Hence, the proof is generally considered pure. On the other hand, a proof of IP generally considered to be impure is the following from (Furstenberg, 1955).

*Impure proof.* The set $\{B_{a,b} \mid a, b \in \mathbb{Z}, b > 0\}$ is a basis for the topology of the integers, containing *arithmetical sequences*, i.e. sets $B_{a,b} = \{a + bn \mid n \in \mathbb{Z}\}$. Now take $B_{0,p} = \{pn \mid n \in \mathbb{Z}\}$. By the Fundamental Theorem of Arithmetic, every integer except $\pm 1$ is contained in some $B_{0,p}$. Let $A = \bigcup_p B_{0,p}$ for $p$ prime, then

| | | |
|---|---|---|
| 7 | 8 | 9 |
| 4 | 5 | 6 |
| 1 | 2 | 3 |

Figure 2.2: A calculator keyboard.

$A = \mathbb{Z} - \{-1, 1\}$. Now it can be shown that $A$ is a union of infinitely many $B_{0,p}$ (and so, that there are infinitely many primes). Namely, if there were only finitely many $B_{0,p}$, this means that $A$ is a closed set; then $\{1, -1\}$ (as $A$'s complement) is open; but each open set must contain a basic open set (one of the arithmetic sequences $B_{a,b}$); this cannot be the case. □

According to Arana and Detlefsen (2011), the proof uses set-theoretical and topological constructions that "clearly [lead] outside the topic" of IP. Specifically, the proof introduces the notions of a topology, arithmetical sequences, notions of 'open' and 'closed' sets, and 'sets' generally, that we do not recognize as belonging to a basic arithmetical topic that IP naturally belongs to. IP does not explicitly mention these notions from set theory and topology, nor does it seem to refer to them implicitly, hence they can be thought of as foreign elements occurring in the proof.

## 2.2 Explanatory value of proof

We turn now to explanation of proof. This virtue of proof starts from the idea that some proofs of a theorem show *why* the theorem is true, and possess explanatory value, while others merely show *that* the theorem is true. A standard example (see e.g. (Lange, 2016b)) concerns a calculator keyboard, see Figure 2.2. Using this keyboard, six-digit numbers can be created by taking any row, column or main diagonal on the keyboard for the first three numbers, and then taking them in reverse order for the last three numbers. It can be proven that any such number (e.g., 789987) is divisible by 37. A proof of this fact that does not possess any explanatory value, simply takes any of our 'calculator numbers', and checks case by case whether this number is indeed divisible by 37. We are now convinced that the theorem is true, but we do not see any reason for it to hold. It still seems just a coincidence, or an accidental truth. However, the following proof (originally given in (Nummela, 1987)) arguably *does* provide an explanation for this truth.

*Proof.* Any calculator number is generated by an arithmetic progression of three integers, which can be written as $a$, $a + d$ and $a + 2d$. Now we can rewrite the entire calculator number as $10^5a + 10^4(a + d) + 10^3(a + 2d) + 10^2(a + d) + a$. Some more rewriting provides the equality of this number to $a(10^5 + 10^4 + 10^3 + 10^2 + 10 + 1) + d(10^4 + x \times 10^3 + 2 \times 10^2 + 10)$, which equals $111111a + 12210d = 1221(91a + 10d) = (3 \times 11 \times 37)(91a + 10d)$. □

The latter product shows that any calculator number is divisible by 37, and thus that the result is indeed not accidental: the more general notation of calculator numbers shows that (among others) 37 unifies all particular instances. In contrast, the case-by-case proof employs 'brute force' that fails to explain.

The distinction between explanatory and unexplanatory proofs, like purity, goes back to ancient Greek philosophers (Lange, 2016b), and a myriad of proposals have been put forth for what it means for a proof to be explanatory. Mathematical explanation is commonly separated from *causal* explanation, where causes explain their effects, since the kind of explaining that mathematical proofs do is taken to be independent of cause or of time. In order to be more precise, models of explanation have been developed, each of which proposes a different conception of the asymmetric explanatory relation between proof and theorem (that does not exist between theorem and proof). While there is no space to discuss all available proposals of explanation, we will briefly describe the most well-known ones in the literature here, and select some of them for the comparison to purity.

First, Marc Lange advertises in various studies (e.g. (Lange, 2014, 2016b)) an approach to explanation that relies on *symmetry* and *salient properties* occurring in the proof. He claims that "a mathematical result that exhibits symmetry of a certain kind is explained by a proof showing how it follows from a similar symmetry in the problem. Each of these symmetries consists of some sort of invariance under a given transformation; the same transformation is involved in both symmetries." This high-level characterization is intended to apply to several known examples of explanatory proofs, but not to exclude explanatory proofs that work differently.

Two of the most well-known early models of explanation are given by Kitcher (1989) and Steiner (1978a). Kitcher claims that a main characteristic of an explanatory proof is its *unificatory power*, while Steiner proposes that a 'characteristic property' of elements occurring in the theorem must play a role in an explanatory proof. More recently, Pincock (2015) proposes an account of explanation that relies on an abstraction difference between proof and theorem. More details about these accounts will follow in the next sections. More accounts of explanation are still being developed, for instance 'conceptual explanation' of Poggiolesi and Genco (2023) where the notion of conceptual complexity is the relevant property.

Recent trends in the study of explanation also include a 'functional' account of explanation Inglis and Mejía-Ramos (2021) (emphasizing that "explanatory criteria offered by earlier accounts can all be thought of as features that make it more likely that a mathematical proof will generate understanding", instead of as

features that 'define' explanation), new case studies (see for instance (Antos and Colyvan, 2024) for a set-theoretic case study) and critical evaluations of established approaches (for instance, Steiner's approach knows criticism from (Resnik and Kushner, 1987) as well as modern literature (D'Alessandro, 2019)), to show where the limits of these accounts lie. Finally, it is interesting to note that the literature on explanation contains several debated 'border' cases, for instance methods of indirect proof (proof by contradiction), and proof by induction. Additionally, there are also strands of explanatory proofs that use specific reasoning strategies, such as explaining by using diagrams (see (D'Alessandro, 2020; Brown, 1999)), or proofs that explain by drawing on analogies (see (Lange, 2016a)). For a broader overview of explanation, we also refer to Mancosu (2001) for a general reflection on explanation, and to (Mancosu et al., 2023) for a recent overview of the literature on explanation.

The main example that we will focus on in our case study, is the following.

### 2.2.1 Pythagoras's Theorem

Consider the following theorem.

**2.2.1 Theorem.** ***Pythagoras's Theorem.*** *For any right triangle $ABC$ with the right angle on $A$, the square $BC^2$ on the hypotenuse $BC$ equals the sum of the squares $AB^2$ and $AC^2$ on the other sides $AB$ and $AC$ of the triangle.*

We present two proofs of this theorem: the first is recognized to have explanatory power, whilst the second is not. Consider the first.

*Explanatory proof.* Take any right triangle $ABC$ with the right angle on $A$ and let $AH$ be its height with respect to the hypotenuse $BC$, as shown in Figure 2.3. First, we notice that the triangle $AHC$ contained in $ABC$ is similar to $ABC$ itself, since they both have a right angle and they both share the angle on $C$.[1] The similarity between the triangles $ABC$ and $AHC$ implies the following equality of ratios:

$$\frac{AC}{BC} = \frac{CH}{AC}$$

which can in its turn be expressed in the following way: $AC^2 = CH \cdot BC$.

Symmetrically, the triangle $ABH$ contained in $ABC$ is similar to $ABC$ itself since they both have a right angle and they both share the angle on $B$. The similarity between the triangles $ABC$ and $ABH$ implies the equality of ratios:

$$\frac{AB}{BC} = \frac{BH}{AB}$$

[1] Since by fixing two angles of a triangle, we also fix the third, the equality of the two right angles of the two triangles and of their angle on $C$ implies that the angle on $B$ of $ABC$ and the angle on $A$ of $AHC$ are equal, which, by the definition of similarity, implies that the two triangles are similar.

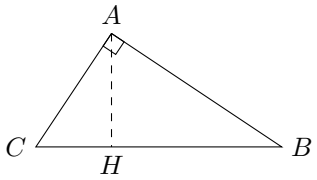



Figure 2.3: A right triangle with height $AH$

which can in its turn be expressed in the following way: $AB^2 = BH \cdot BC$. Since we have established that $AC^2 = CH \cdot BC$ and $AB^2 = BH \cdot BC$, we can put them together, obtaining $AC^2 + AB^2 = CH \cdot BC + BH \cdot BC$, which is equivalent to $AC^2 + AB^2 = (CH + BH)(BC)$. This, in turn, precisely gives us the conclusion $AC^2 + AB^2 = BC^2$, since by construction $CH + BH = BC$. □

The first mathematician who argued that this proof is explanatory was Bouligand (1937, p. 258). Both Bouligand and Steiner (1978a) (see also (Mancosu, 2001)) claim that the explanatory character of the proof is due to its general nature. The proof explains a property of right-angled triangles by relying on a more general property which belongs to 'similar' figures: the latter emerges as the reason why the former is true. The explanatory nature of this proof is arguably less universally agreed upon than, for instance, the example concerning the calculator keyboard. This illustrates already the level of open texture that ideals of mathematical proof may contain. Hence, if desired, a more refined starting point for this proof is to take it as embodying a more specific *variant* of explanation, for instance as the type of explanation that suitably matches the sharpened concepts of Steiner (1978a)'s model (which we will describe in more detail in Section 2.5).

For a contrasting unexplanatory proof of Pythagoras's Theorem, here is one which serves the purpose.

*Unexplanatory proof.* Consider any right-angled triangle $ABC$, where the hypotenuse $BC$ is called $c$ and other sides $AB$ and $BC$ are called $a, b$, respectively. Consider then the square $q$ with side $a + b$ constructed by four copies of the triangle $ABC$ as shown in Figure 2.4. Clearly, the area of the internal square $q'$ with side $c$ is equal to the difference between the area of $q$ and the area of the four copies of the right-angled triangle, that is, $Area(q') = Area(q) - 4(Area(ABC))$. Now, we know that the area of $q'$ is $c^2$, the area of $q$ is $(a + b)^2$ and the area of one triangle is $\frac{ab}{2}$, therefore we have that $c^2 = (a+b)^2 - 4(\frac{ab}{2})$. By simplifying the product with the fraction, we obtain the equality $c^2 = (a + b)^2 - 2ab$ and then, by rewriting the square of the sum, we obtain the equality $c^2 = a^2 + b^2 + 2ab - 2ab$ which results in the desired equality: $c^2 = a^2 + b^2$, namely $BC^2 = AB^2 + BC^2$. □

By the constructions carried out and subsequent reasoning about the surface

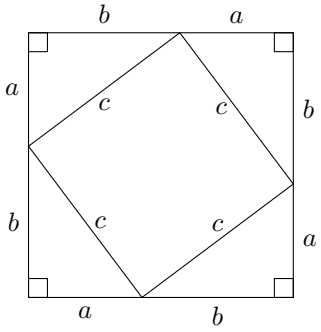



Figure 2.4: Arrangement of four copies of a right-angled triangle around a square.

areas, we become convinced that Pythagoras's Theorem holds. However, it remains unclear why we made the particular arrangement of Figure 2.4 in the first place, as it does not lead to any general property the theorem eventually relies on. In other words, there is no element of the proof that seems to stand out as the reason why the theorem is true, hence, the proof is (intuitively) unexplanatory. Consider Bouligand's words, that also apply to this example:

> "One sees clearly the difference between these two ways of treating the same problem. Only the *first* gives a satisfactory explanation, precisely because it takes place within the domain of causality of the propositions that is to be established" (Bouligand, 1937, p.7).

## 2.3 Comparing purity and explanation

Below we sketch possible similarities between purity and explanation through a historical conception that connects them through an 'order of truths'. Although this should not necessarily be interpreted as a universally held perspective on purity and explanation at the time, it was to a certain extent supported by the scholars mentioned below. It emphasizes an interesting harmonic view on these two ideals, that suits comparisons to their (more diverging) formalized models that show where sharpenings depart from this pre-theory.

One of the first to establish a link between pure and explanatory proofs was Aristotle. In particular, Aristotle underlines the epistemic value of pure proofs. To him, a pure proof shows that the predicate of its conclusion holds of its subject solely because of what the subject in itself is (see Detlefsen (2008)). It follows that a pure proof reveals the reason why the subject has the property expressed by its predicate, since reasons are typically extrapolated from the very essence of the subjects themselves.

> The 'why' is referred ultimately [...] in mathematics, to the 'what', (to the definition of 'straight line' or 'commensurable', &c.) [...] (Aristotle, 1954, §198)

Because pure proofs provide the reasons why their conclusions are true, they also become explanatory proofs. This conception of proofs, that are at the same time explanatory and pure, can also be found in the work of Leibniz. According to Leibniz, there is an ordering among mathematical truths, and explanatory proofs reveal this order by showing the reasons — namely the more fundamental truths — why the theorem is true. However, this objective ordering of truths also suggests a conception of purity: an explanatory proof is also a pure proof in that it is restricted to a segment of this ordering of truths.

> [W]e are not concerned here with the sequence of our discoveries, which differs from one man to another, but with the connection and natural order of truths, which is always the same. (Leibniz, 1981, Bk. IV, ch. vii, §9)

The dichotomy of explanatory proofs and pure proofs finds support in the work of Bernard Bolzano. Bolzano distinguishes between proofs which merely show *that* a theorem is true, and proofs which explain *why* a theorem is true. Thales, Bolzano (1837) says, did not attempt to prove that the angles at the base of an isosceles triangle are equal, though this was evident to him. Rather, he tried to find the reason for this truth: he tried to provide an explanatory proof of this theorem. The reasons for a certain truth are for Bolzano always simpler than the given truth. In other words, for any given truth, if we aim to find its grounds, we need to look inside the truth itself, to analyze the constituents of its concepts. In this respect, a proof which explains by providing the reasons why its conclusion is true, is also a pure proof, in that the reasons are always connected to the conclusion and are simpler, or more general[2], than the conclusion itself.

> [It is] an intolerable offense against correct method to derive truths of pure (or general) mathematics (i.e. arithmetic, algebra, analysis) from considerations which belong to a merely applied (or special) part, namely, geometry [...] if one considers that the proofs of a science should not be merely *confirmations* (*Gewissmachungen*), but rather *justifications* (*Begründugen*), i.e. presentation of the objective reason for the truth concerned, then it is self-evident that the strictly scientific proof, or the objective reason, of a truth which holds for *all* quantities, whether in space or not, cannot possibly lie in a truth which holds merely for quantities which are in *space*. (Russ, 1980, p. 160, translating Bolzano)

[2]As underlined in several texts, e.g. Ginammi et al. (2020) or Poggiolesi (2024), complexity and generality are often linked; more precisely, they are inversely proportional.

A last advocate of the connection between explanatory and pure proofs is Frege (see Detlefsen (2008)). Not only did Frege accept purity as an ideal of proof, he also connected it with proofs that explain by the idea that pure proofs provide the reasons for the truth of a theorem.

> The aim of proof is in fact not merely to put a proposition beyond all doubt, but also to afford us insight into dependence of truths upon one another. [...] The further we pursue these enquiries, the fewer become the primitive truths to which we reduce everything, and this simplification is in itself a goal worth pursuing. (Frege, 1980, §2)

Despite this historical tradition, that from now on we might call the *Aristotelian tradition*, the connections between explanation and purity have loosened in the contemporary literature. On the one hand, the notion of explanatory mathematical proof has increasingly encompassed cases of mathematical proofs that, instead of explaining by providing the reasons why the theorem is true, rather explain by displaying a diagram, or by drawing on an analogy, see for instance (D'Alessandro, 2020). However, once an explanatory proof does not necessarily provide the reasons, which should be simpler and thus pure (according to the Aristotelian tradition), of why its conclusion is true, its connection with purity is undoubtedly blurred. On the other hand, although the notion of purity still concerns itself with certain restrictions in the way we are allowed to prove a theorem, it has been given several new interpretations, which part from the idea of the ordering of truths that was at the core of the Aristotelian tradition. Once the connection to the hierarchy of truths is lost, the relationship with the traditional notion of explanation, where reasons are simpler than the conclusion, fades away as well. As a result, nowadays more and more scholars insist on the importance of looking at the properties of explanatory power and purity of proof as separate. Whilst Lange (2015), for example, when analysing the proof of Desargues's theorem in projective geometry, concludes that such a proof shows that "a proof's explanatory power is independent of its purity"[3], Lehet (2021) or Ryan (2023) more generally aim at showing that impurity, rather than purity, might extend our knowledge and thus provide explanation.

### 2.3.1 A new comparison

In view of the diversity of these positions, we aim to develop a study which analyzes the links between explanation and purity of mathematical proofs, i.e. which shows if and where the two notions coincide. The comparison of ideals of proof is of general philosophical interest, as shown for instance by Pel (2023) (comparing purity to 'directness' of proof), Arana (2017); Iemhoff (2017) (comparing

[3] Although we do not discuss Desargues's Theorem in this chapter, it is fairly common in the literature on purity — see also (Arana and Mancosu, 2012).

(im)purity to simplicity of proof), and Novaes (2019); Lange (2016b); Inglis and Aberdein (2015) (relating 'beauty' of proofs to other ideals of proof, such as explanation and simplicity). Such investigations allow us to evaluate our intuitions concerning these ideals more critically, and to better understand the epistemic differences between various proofs of the same theorem (see also Dawson (2006)). The comparison between models of purity and explanation of proof has been initiated by Arana (2022), who argues that 'topical purity' as in (Arana and Detlefsen, 2011) and the notion of explanation of Steiner (1978a) do not coincide. We here expand this study by providing a comparison with the following features.

1. We take into account a *wide* variety of models of purity and explanation.

2. We take a *case study* of two mathematical theorems — Pythagoras's Theorem and the Infinitude of Primes — and their proofs.

3. For the comparison to be clearly structured, we will rely on the distinction between *epistemic* and *ontic* models.

First, treating a wider variety of models of purity and explanation allows for a more nuanced comprehension of the interaction between purity and explanation. Here, each model can be seen as a unique perspective of what it means for a proof to be pure or explanatory, while we do not impose on any individual model to be absolutely representative of explanation or purity as a whole (in the spirit of (Inglis and Mejía-Ramos, 2021)). That is, we cannot claim to make a completely general comparison between explanation and purity as single ideals of proof. Instead, we are essentially contrasting a given proof with the various *aspects* of purity and explanation that are represented by different models. We can only make tentative conclusions about a generic interaction between purity and explanation, for instance in case that different models turn out to have similar judgements on the same proof.

Second, in order to feasibly compare the variety of models, we restrict to a case study of Pythagoras's Theorem and the Infinitude of Primes. We justify their choice by their existence in the literature on purity and explanation, respectively.[4] Moreover, these examples are relatively simple and thus naturally allow for a cross-model comparison. For a more complete picture, of course, more contemporary examples from different areas of mathematics (in the spirit of studies like (Colyvan et al., 2018)) should be incorporated in the comparison. Our analysis is then 'systematic' by fixing a proof under scrutiny, and by methodically applying it to different models of purity and explanation from the literature. In particular, a

[4]These proofs have received a good amount of consensus; however, note that neither in the literature on explanation, nor in that on purity, consensus about which proofs are explanatory or pure is universal. Additionally, note that all of our models of purity consider the relevant proof of the Infinitude of Primes pure, and at least Steiner's model of explanation considers the relevant proof of Pythagoras's Theorem explanatory. Skeptics of the proofs in our case study will at least concede that we are comparing various aspects of *models* of explanation and purity.

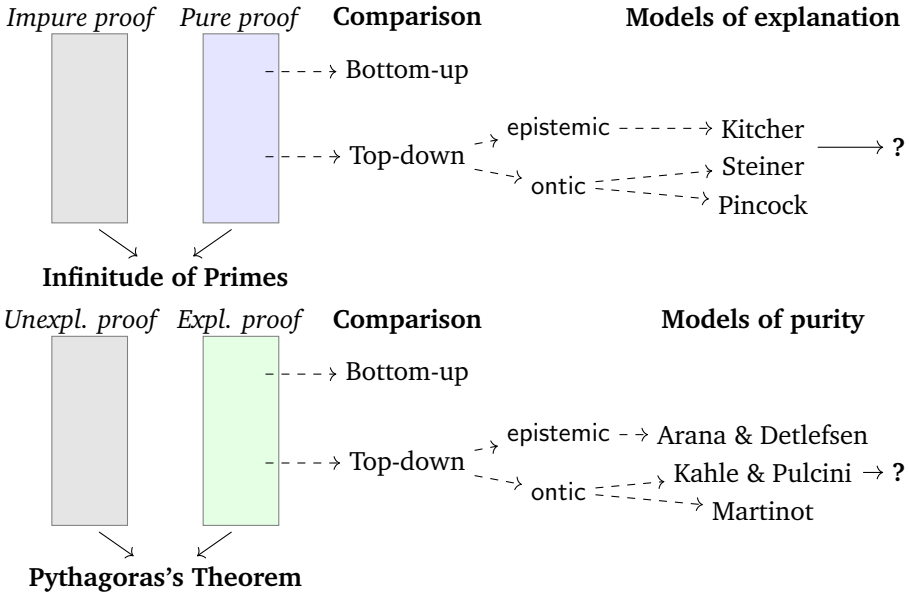



Figure 2.5: A visualization of the comparison between purity and explanation.

proof of the Infinitude of Primes that is generally considered pure will be analyzed on its explanatory value according to a variety of models of explanation, and a proof of Pythagoras's Theorem that is generally considered explanatory will be analyzed on its purity according to a variety of models of purity.

Note that we do thus not aim to make a comparison about *all* pure and explanatory mathematical proofs. By taking a case study of two particular proofs, "we run a risk of developing an account of explanation [and purity] that is based on too limited a stock of examples and does not do justice to mathematics as a whole" (Antos and Colyvan, 2024). However, our current focus is to maintain a diverse view of explanation and purity by the multitude of models, while studying proofs that are likely to produce relatively stable results under these models.

Third, for the comparison to be clearly structured, we will rely on two well-known distinctions which emerge in the literature on explanation as well as on purity (see Figure 2.5 for an overview of the approach). The first distinction is that between bottom-up and top-down approaches. The bottom-up perspective begins by avoiding, as much as possible, any commitment to a particular theoretical framework, and mainly works from the judgements of mathematicians in practice. These judgements decide which proofs are called (im)pure or (un)explanatory (based on a primitive understanding of the notions). Only afterwards a taxonomy of recurrent types of explanatory mathematical proofs and pure mathematical proofs is provided, which allows one to investigate whether these patterns are

heterogeneous, or whether they can be subsumed under a general account. We will omit a discussion of a bottom-up comparison between purity and explanation, as discussions in the literature are relatively scarce and do not add much to the current comparison — instead, we focus on the top-down approach of the chosen models. A top-down characterization of the notions of explanation and purity follows theoretical intuitions about their nature. Their intuitive properties are made precise in a theoretical framework, often resulting in a model which accounts for the purity and explanatory power of as many mathematical proofs as possible.

The second distinction classifies the top-down models in a more refined way. This concerns the difference between *internalist* or *epistemic* models, and *externalist* or *ontic* models (see Kim (1994)). Whereas epistemic models are models which account for the purity or the explanatory power of mathematical proofs by looking at proofs as activities internal to an epistemic corpus (a theory or a set of beliefs), ontic models look for some systematic pattern of objective dependence relations which explanatory or pure proofs can track or can be identified with. In particular, the most notable epistemic accounts of purity and explanation are those of Arana and Detlefsen (2011) and Kitcher (1989), respectively; whilst the most notable ontic account of explanation is given by Steiner (1978a), and a more recent one is that of Pincock (2015). Two ontic approaches to purity have more recently emerged, namely those of Kahle and Pulcini (2017); Martinot (2024a). We thus combine more traditional as well as more contemporary perspectives, taking a broad and tolerant view of what aspects can intuitively underlie purity and explanation. As we will see, although for most of the models proposed in the recent literature, purity and explanation are two separate features of proofs, there exists a recent account of explanation where the two notions may be seen as reunited again.

Finally, note that even restricted to our case studies, and to individual models, there are instances where our results are open to interpretation. For instance, occasionally, the precise wording of a theorem can affect its outcome in analyses of explanation, but especially of purity. Specifically, this occurs in the notion of 'topic' in (Arana and Detlefsen, 2011) (see Section 2.4.1), and in the notion of 'ontology' of (Martinot, 2024a) (see Section 2.4.1). Where this comes into play, we aim to objectively discuss the different formulations, and we will consider potential differences in results.

In what follows, Section 2.4 will be dedicated to describing three models of purity and their application to the explanatory proof of Pythagoras's Theorem. Section 2.5 will present three models of explanation, and consider their application to the pure proof of the Infinitude of Primes. Section 2.6 will evaluate some aspects of our models and comparison.

## 2.4 Models of purity

We will zoom in on the theoretical ingredients of the following models of purity:

Arana and Detlefsen (2011)'s *topical purity* (an epistemic model)

Kahle and Pulcini (2017)'s *operational purity* (an ontic model)

Our model of *ontological purity* (an ontic model, see (Martinot, 2024a))

These models all provide an individual take on what elements a pure proof should restrict itself to, appealing to notions such as our understanding of a theorem, the operational strength of a theorem, and various ontologies that a theorem can be said to refer to.

**Arana and Detlefsen's model of purity**

Arana and Detlefsen (2011)'s *topical purity* aims to identify pure proofs via the epistemic values that they possess: pure proofs relieve an agent of a particular type of ignorance she might have. Generally, topical purity explicates the relation between proof and theorem as a problem in a yes-no form that an investigator attempts to solve, the result of which needs to relieve her ignorance.

More precisely, a *directed problem* $\mathcal{P}$ is a triplet $\mathcal{P} = (?_{y/n}, P, \phi)$, which consists of a yes-no interrogative attitude, a propositional content $P$, and a formulation of $P$, namely $\phi$, in terms of a language or theory that is available to the investigator. Relieving ignorance of $P$ can be done in two ways: either by *solving* it and providing an answer in yes- or no-terms, or by *dissolving* the question. The latter means that the problem ceases to be a problem for the investigator, because one of the commitments made in representing it is retracted. That is, let $\mathcal{E}$ be the solution of the investigation of the problem. Then the investigator may at some point stop accepting some premise or inference in $\mathcal{E}$, one which may happen to be important in representing the problem itself as well — so that once this premise or inference is removed, the entire problem is dissolved. The idea is thus that solutions $\mathcal{E}$ can *share commitments* with the problem $P$. Retracting shared commitments can change the problem $P$, and take away its need to be solved.

Then a *co-final* solution $\mathcal{E}$ to the problem is one where solving and dissolving the problem are strongly linked: $\mathcal{E}$ either survives as a *solution* to the problem, or, if at any point it *stops* being a solution to the problem (because of some retraction of a commitment), then the problem itself is dissolved, and also stops being a problem. That is, there is no situation where a retraction of a commitment that affects the solution, does not affect the problem itself: if $\mathcal{E}$ does not solve the problem, it will delete the problem. Furthermore, an investigation *stably solves* its problem $\mathcal{P} = (?_{y/n}, P, \Phi)$ when it offers a solution $\mathcal{E}$, which provides both evidence justifying belief in a yes- or no-answer to the problem (i.e., it provides a solution, and not (yet) a dissolution), and has the property of *co-finality*: it is the

type of solution where any retraction of premises or inferences in $\mathcal{E}$ would *dissolve* the problem for the investigator.

The connection to purity of proofs is finally made through co-finality. Namely, in general, a proof of a theorem is pure if its resources are restricted to those that 'determine the content of a problem'. There are different ways to spell out 'determine', and co-finality provides one of them. That is, one can restrict a proof exactly to the set of commitments such that, if they are retracted by the investigator, the content of the problem changes for her (the commitments that ensure co-finality). Arana and Detlefsen (2011) call the *topic* of the problem exactly this set of commitments. To give a better image of these commitments, they cite as examples "definitions, axioms concerning primitive terms, inferences, etc". Then a solution $\mathcal{E}$ of $\mathcal{P}$ is *topically pure* when it draws only on such commitments as topically determine $\mathcal{P}$. Topical purity retains its epistemic significance by providing stable (and so co-final) problem solutions. A topically impure solution, on the contrary, would contain commitments that can be retracted *without* dissolving the problem.[5]

**2.4.1 Example.** An understanding of the Infinitude of Primes (IP) 'at face value' is given by Arana and Detlefsen (2011) as several commitments: axioms for successor and induction, definitions and axioms for an ordering of the natural numbers, a conception of primality and definitions of divisibility and multiplication. A natural formulation of these commitments takes place in first-order Peano Arithmetic. The pure proof in Section 2.1.1 can naturally be carried out in PA, so that the commitments just mentioned are a suitable *topic* of IP. Retraction of any of the commitments in the topic, according to Arana and Detlefsen (2011) requires a corresponding change in our understanding of IP, and the proof acquires topical purity.

### Kahle and Pulcini's model of purity

Kahle and Pulcini (2017) aim to provide a notion of purity that focuses on the ontology of mathematical operations and "avoid[s] any reference to a knowing subject", and so avoids epistemic factors. Additionally, they insist that when a proof is impure, it should not be that "the resort to [...] extraneous notions is nothing else but an avoidable roundabout". That is, for a proof to be impure, the notions that are extraneous in it must really be essential to the proof — otherwise a proof was actually pure after all. These desiderata require a different measurement of 'extraneousness' than, for instance, the epistemic notion of Arana and Detlefsen (2011), that is based on the understanding of a given investigator.

To start, the *operational content* of a theorem $T$ or a proof $P$, denoted by $C(T)$

[5]Note that in this context Arana treats topic determination in a naive way "as doing so is consistent with the way mathematicians have treated purity in practice" (Arana, 2022). This means that it is in principle dependent on an agent how to understand the theorem.

and $C(P)$, respectively, is given by a set that contains the mathematical operations mentioned in the theorem or in the proof. Given these sets, their *ontology*, $D(C(T))$ and $D(C(P))$, are the smallest numerical domains ($\mathbb{N}, \mathbb{Z}, \mathbb{Q}, ...$) that are closed under the operations in $C(T)$ and $C(P)$. For two operational contents $C_1(X_1)$ and $C_2(X_2)$, one then writes $C_1(X_1) \preccurlyeq C_2(X_2)$ in case $D(C_1(X_1)) \subseteq D(C_2(X_2))$. Finally, a proof $P$ of a theorem $T$ is *operationally pure* if the ontology of the proof is a subset (even an improper one) of the ontology of the theorem, i.e. if $C(P) \preccurlyeq C(T)$.

**2.4.2 Example.** As an example of the operational content of IP, and of Furstenberg's proof of IP ($P_{\mathsf{IP}}$), Kahle and Pulcini (2017) suggest that $C(\mathsf{IP}) = \{\backslash\}$, as "[t]he definition of prime number relies on the division operator". This means that, as the smallest set closed under the arithmetical operation of division, we have $D(C(\mathsf{IP}) = \mathbb{Q}$. Though not mentioned in their work, they undoubtedly evaluate Euclid's proof of IP as operationally pure, remaining within the numerical domain $\mathbb{Q}$. Interestingly, they also consider Furstenberg's proof operationally pure, as "the inclusion of union and intersection in any operational content do not affect the underlying ontology", which remains $\mathbb{Q}$.

Generally speaking, however, there may be a certain arbitrariness in practice in selecting the operations 'mentioned' by a theorem or proof, as it is unclear when exactly this mentioning occurs, and what is the distinguishing feature of operations. Still, in what follows, we will attempt to apply the framework of operational purity consistently to the explanatory proof of Pythagoras's Theorem.

### Our model of purity

We have introduced 'ontological purity' in (Martinot, 2024a), which will be elaborated on more in Chapter 3 and 4. In order to apply it in this chapter, we provide a brief summary of it here. Our model introduces two levels of purity: full ontological purity and secondary ontological purity.[6] A central idea for this type of purity is that a mathematical theorem is about "those things to which the terms appearing in it refer" (Detlefsen, 2008). More specifically, to each theorem one should associate an ontology of mathematical objects and operations that it is primarily intended to talk about. Intuitively, if a proof only makes use of elements of this ontology, it is ontologically pure; but additionally, if it makes use of a different ontology, yet restricts itself to 'simulations' of elements of the pure ontology, it can still obtain a secondary level of ontological purity.

To make this more precise, given a paired-up theorem and ontology, the approach asks one to additionally select a first-order mathematical '*context theory*' $T$ that is (according to the interpreter of this theory) seen to best capture this ontol-

[6] While a large part of Chapters 3 and 4 consider the purity of natural deduction proofs, we here only focus on its discussion of 'informal' proofs.

ogy. An ontology can be thought of as similar in nature to an intended standard model of this theory. A standard example would be to consider the ontology of natural numbers together with the usual arithmetical operations, and pick Peano Arithmetic (PA) as the context theory referring to this ontology. While the signature of the context theory (the primitive predicate symbols, function symbols and constants) can be seen to pick out basic ontological objects and operations, the axioms of the theory determine the full size and complexity of the ontology. What kinds of intuitive objects exactly make up this ontology, however, is ultimately decided by the individual interpreting the signature of a theory (for instance, PA may also be considered to primarily capture binary strings, instead of numbers).

For full ontological purity, then, we need to find out if each notion occurring in the proof belongs to the ontology of the theorem. Given that the context theory captures this ontology, we can do this by judging whether there exists a '*natural formalization*' of any notion in the proof into the context theory. That is, if each notion in the proof has a definition in terms of the primitives of the context theory, we can check whether this definition makes ontological sense, and whether it is relatively elegant and efficient, supporting the naturalness of its formalization. For instance, while the primitives of PA pick out principal ontological elements that allow easy definitions of notions like primality, the axioms of ZFC focus on main properties of sets and require building up arithmetical notions from scratch. The judgement of full ontological purity, however, remains inherently subjective, so that its results are always relative to preferences of the mathematical community, and to changing views over time.

For secondary ontological purity, we need to find out if a proof in some sense restricts itself to 'simulations' of elements of the ontology of the theorem, and this requires a slightly more intricate approach. Intuitively, we want to judge whether each notion occurring in the proof is 'just a disguise' of an element from the ontology of the theorem. For this, several requirements should be satisfied. First, the notions in the proof should have a natural formalization in a theory $V$ that is *different* from $T$ — meaning that $V$ captures a different ontology than that of the context theory (although it could be that differences are small). It is not always easy to judge whether these formalizations in $V$ are just simulations of pure ontological elements, however, as there can be a lot of different types of simulations. For instance, a proof in set theory that only deals with the Von Neumann ordinals intuitively restricts itself to just simulations of the ontology of Peano Arithmetic (natural numbers); but there may also exist less natural codings of natural numbers that are harder to recognize as 'just number simulations'. Hence, to check whether the proof can be seen as restricted to simulations, even if our intuitions abandon us, there are two more requirements to check.

For one, we should check whether any notion in the proof is (in any way) formalizable in the context theory $T$. We leave it to mathematicians in practice to decide what counts as a (possibly unnatural) formalization. For example, a theory like PA cannot always deal with big infinities, but may be able to represent

them with finite elements: one may wonder whether this is still a formalization of such infinities. Second, however, once we know that any notion of the proof is formalizable in $T$, we just need to know that $V$ can simulate these formalizations. For this, we implement the formal requirement of checking whether there exists an interpretation between them. An interpretation of $T$ into $V$ ($i : T \to V$) as defined by Visser (1997) amounts to a translation $i$ of $T$-predicates, function symbols and constants to $V$-formulas, and to a formula $\delta$ describing a restriction of the domain that $V$ talks about to the objects that are able to 'simulate' objects from $T$ (we call such objects 'surrogates'). The interpretation result then says that restricted to this domain, $V$ can prove all translated $T$-theorems. We argue that an ontology can be seen to have an underlying *ante rem structure* (as in (Shapiro, 1997)), and that this structure is preserved under interpretations. This justifies more the secondary purity result, by claiming that the ontology of the context theory and surrogate ontology have structural content in common.

Hence, if a proof satisfies all three conditions, it is secondarily ontologically pure with respect to $V$. Intuitively, secondary ontological purity means that, although a proof primarily refers to an ontology that can have a fundamentally extraneous nature, this domain can be restricted to *just* representations of the fully pure ontology of the theorem (the surrogate ontology). If, instead, a proof makes use of a notion in $V$ that simply cannot be represented by the 'pure $T$-surrogates' in $V$, then extraneousness is counted more heavily, and the impurity result is more severe. For an example of full ontological purity, we refer to Example 3.4.4 of Chapter 3.

### 2.4.1 Application to the explanatory proof of Pythagoras's Theorem

Equipped with the tools of three models of purity, this subsection will see them in action, leading to the following results.

> The explanatory proof of Pythagoras's Theorem is *impure* according to Arana and Detlefsen (2011)'s *topical purity*.
>
> The explanatory proof of Pythagoras's Theorem is *impure* according to Kahle and Pulcini (2017)'s *operational purity*.
>
> The explanatory proof of Pythagoras's Theorem can be interpreted as *pure* as well as *impure* according to our model of *ontological purity*.

#### Arana and Detlefsen's model of purity

Although a first analysis is already given in Arana (2022), arguing that the explanatory proof of Pythagoras's Theorem is topically impure, we here attempt to expand on it more, and we attempt to take a tolerant attitude in following the

method of topical purity as described in (Arana and Detlefsen, 2011) (taking inspiration from Arana and Detlefsen's discussion there of Sylvester's problem and Furstenberg's proof of IP).

Consider Pythagoras's Theorem together with its explanatory proof, as introduced in Section 2.2.1. Following the details of Arana and Detlefsen's account given above, Pythagoras's Theorem can be restated as the problem of whether it holds that for any right-angled triangle with sides $A, B, C$ and right angle $\angle AB$, $A^2 + B^2 = C^2$. It seems reasonable to say that the topically determining commitments of this problem (i.e. what the content of the problem is) include the definitions or axioms for the concepts of points, lines, (right) angles, degrees, length, triangles, squares, surface area; and arithmetical definitions or axioms for numbers, addition, equality, multiplication and exponentiation (by a factor two). These elements are all directly involved in the statement of the problem (i.e. Pythagoras's Theorem) and to our knowledge they are all that is needed to fully understand the problem itself in the intended basic and direct way.

Consider now the explanatory proof of Pythagoras's Theorem. This proof crucially relies on the fact that a right-angled triangle can be split up by its height into two triangles that are similar to it, and to each other. Then by the fact that the sides of all these triangles are proportional, we know that the equality $A^2 + B^2 = C^2$ follows. In other words, our proof or investigation relies on a notion of similarity. Similarity can be made precise in two different (but equivalent, in plane Euclidean geometry) ways:

> *Formulation 1.* Two triangles are *similar* if, and only if, all corresponding angles have the same measure.
>
> *Formulation 2.* Two triangles are *similar* if, and only if, the lengths of their corresponding sides are proportional.

Now the proof would be co-final with the problem if any commitment in the proof is shared by the problem (the topic), so that if you retract it, the content of the problem changes. However, we aim to defend the idea that the notion of similarity (in whatever formulation) violates this property. That is, we could 'retract' our commitment to both formulations of similarity, without changing our understanding of Pythagoras's problem.

To see this, note that depending on the particular theorem, it can matter whether one or *all* definitions of a concept occurring in the proof are retracted. Arana and Detlefsen (2011) note that in principle, retracting one formulation of a concept breaks the connection between said formulation and the concept it defines, but it does not prevent this concept from still representing something to us via some other definition. For instance, consider Sylvester's problem, described in (Arana and Detlefsen, 2011), which concerns a geometrical theorem and its proof containing the notion of distance. In that case, a retracted metrical definition of distance is still allowed to be replaced by an order-theoretic definition of distance

instead, and unlike the metrical definition, this other definition does not disrupt the purity result. Here, it matters which definition is retracted: only some formulations prevent dissolution of the problem when they are retracted, and are impure. In other cases, however, any formulation of the concept occurring in the proof is impure, as retracting any of them does not lead to dissolution of the problem. For instance, Arana and Detlefsen (2011) note that Furstenberg's proof of IP "does not require th[e] concept [of topological space] to be represented to [them]sel[ves]" at all, and so it is clear that retracting any definition for topological space does not dissolve the problem. For Pythagoras's Theorem, we argue in the spirit of Arana (2022) that, like the definition of topological space, the concept of similarity does not need to be represented to us at all. Namely, Pythagoras's Theorem does not require us to understand or know about any representation of the notion of similarity. Hence, we take retracting similarity to mean retracting both Formulation 1 and Formulation 2, as we intend to retract the property of *comparing* angles or sides (whichever one), and in fact any knowledge of the relative size of angles or sides. These properties are simply not relevant to the problem: we do not need to know that the angles of several pairs of triangles are exactly the same, or that their sides are proportional, to understand how the surface areas of the squared edges of one of them correlate. Thus, the retraction of the concept of similarity does not lead to the dissolution of the problem. It follows that the similarity proof is not co-final, that it is therefore not a stable solution, and is indeed impure.

We emphasize that the specific ingredients of the definitions of similarity (angles, sides, and so on) are still topically relevant to the problem, and retracting those definitions would be harmful for purity. It is the higher-level property acting on these basic ingredients that we call similarity, and that we retract. This is analogous to Sylvester's problem and the retraction of the metrical definition of distance, which says that there exists a shortest line between every line and point not on that line. In order to understand Sylvester's problem (i.e. that if for $n$ points, any straight line joining two of them passes through a third, then all $n$ points lie on a straight line), it does not matter whether there exists a shortest line between every two points. We just need to understand that a line can intersect with points, without needing to think about how far apart the points are, or whether this line measures that. Thus, the property of distance is in a sense more 'high-level' than needed for our basic understanding of Sylvester's problem, and the same holds for similarity relative to Pythagoras's Theorem.

**Kahle and Pulcini's model of purity**

In order to apply Kahle and Pulcini (2017)'s model of purity to the explanatory proof of Pythagoras's Theorem, we need to select the set of operations mentioned by the theorem and by the proof. While some interesting observations come up in doing this, they are mostly consequences of the practical ambiguities of the model, and we consider our purity evaluation to be subject to potential changes

to the model that resolve these matters.

A main observation is that it seems up for discussion exactly which operations are mentioned by the explanatory proof of Pythagoras's Theorem. A reasonable set of operations mentioned by the theorem could be $\{x+y, x\times y, x^2\}$ (written as how the operations act on number variables $x, y$, for clarity).[7] For the operational content of the explanatory proof of Pythagoras's Theorem, we are first confronted with the issue of how to incorporate geometrical constructions. For instance, the proof involves constructing a line through the right-angled triangle, establishing the similarity result between the three resulting triangles. It is unclear whether the height construction in the main right triangle, and thus the creation of the other two triangles together with the establishment of the similarity between them, involves any operation that should be included in the operational content of the proof. Kahle and Pulcini (2017)'s approach has a strong emphasis on arithmetical operations, which culminates in the measure of operational strength in terms of numerical domains. As a result, even if the above operations were included in the operational content of the proof, we would still lack a direct way of evaluating their strength in terms of a numerical domain.[8] As a consequence, it seems that for now, the best choice is to focus on the arithmetical operations mentioned in the proof — while the question of how operational purity should behave in non-arithmetical mathematical fields remains open, but is something that could be spelled out by generalizations of the model.

Second, even restricting to arithmetical operations, the question arises what exactly is an 'individual operation'. For instance, Kahle and Pulcini consider the operation of division to be essential to the content of the proof. But operations can have duals, that allow rewriting in terms of one another: a division $\frac{a}{b} = c$ can always be written as a multiplication $a = b\times c$, something the similarity proof often uses. That is, although the proof starts from equalities $\frac{a}{b} = \frac{c}{d}$, it would arguably be the same to represent these immediately as equalities $a \times d = c \times b$ concerning multiplications (instead of only as a second step). That is, the operation of division itself may be something that is 'an avoidable roundabout', and so perhaps not really essential to the proof.

This leads to two possible sets of operations, $C_1(P_{\text{PT}}) = \{x+y, x\times y, x^2, x\backslash y\}$ and $C_2(P_{\text{PT}}) = \{x+y, x\times y, x^2\}$. This difference in choice of operational content is decisive for the purity result, as $D(C_2(P_{\text{PT}})) = D(C(\text{PT})) = \mathbb{N}$, but $D(C_1(P_{\text{PT}})) = \mathbb{Q} \supset D(C(\text{PT}))$. As Kahle and Pulcini (2017) note, by adding the division operator to the natural numbers we gain the positive fractions. Note that if we allowed exponentiation generally, instead of just exponentiation by a factor two, we would even obtain negative and irrational numbers, by defining, for instance, $2^{\frac{1}{2}}$, and

[7] Addition is explicitly mentioned by the theorem, and so is exponentiation by a factor two. Since the usual definition of exponentiation relies on multiplication, we consider the latter to be part of the operational content as well.

[8] This is also a reason to consider Pythagoras's Theorem in its arithmetical formulation, instead of a geometrical one, as considered in the next section.

even imaginary numbers (consider $(-1)^{\frac{1}{2}}$). But even with restricted exponentiation, $C_1(P_{\text{PT}})$ is a more powerful combination of operations, giving us the domain of the positive rationals, $\mathbb{Q}^+$.

We can either accept the fact that 'inverse' operations have different operational ontologies, or we might want them to produce the same ontology. The first option sounds reasonable, considering that our understanding of proportionality in terms of division (an act of separation) is arguably different from proportionality in terms of multiplication (an act of combination). However, to allow inverse operations to induce a different purity result does not always seem reasonable. Similar to multiplication and division, consider the case of addition and subtraction: a theorem that only mentions addition will have ontology $\mathbb{N}$. But if its proof mentions subtraction, suddenly we are given the ontology $\mathbb{Z}$. Yet does anyone really believe that subtraction should be an impure element in a proof of an additive statement? This seems a far-fetched claim. Still, the question of what exactly is the difference between inverse or dual operations, remains philosophically interesting. They act on elements in a different order, and this order can make a difference for closure results under a given numerical domain. Whether these matters merely relate to mathematical properties of operations, or also underlie philosophical intuitions about operational strength, is to be determined.

Thus, although we concede with our application that the explanatory proof of Pythagoras's Theorem under Kahle and Pulcini's model should be considered impure here, as it aligns with the way they intend their model to work, we recognize that this result is relative to certain ambiguities.

**Our model of purity**

Now consider the model of ontological purity. It seems quite clear that the ontology of Pythagoras's Theorem should include two-dimensional geometrical shapes, lines, points, and so on. This suggests we should take some geometrical theory as a context theory, such as Euclidean plane geometry, or Hilbert's or Tarski's axioms for plane geometry. At the same time, we might think that the ontology should contain some arithmetic. We suggest that this depends on the specific version one takes of Pythagoras's Theorem: consider two different formulations.

1. *Arithmetical formulation.* For the length of the sides of right-angled triangles, where $A$ is the length of the hypotenuse, it holds that $A^2 = B^2 + C^2$.

2. *Geometrical formulation.* In right-angled triangles, the square on the side opposite the right angle equals the sum of the squares on the sides containing the right angle.

According to the first formulation, the investigator of purity is expected to envision an ontology including the geometrical plane as well as an arithmetical domain. The second formulation only requires an ontology of the geometrical plane. This

affects the choice of context theory. The first formulation could be compatible with a first-order theory axiomatizing both plane geometry as well as arithmetic, or one might be content with a geometrical theory that is able to interpret and define numbers in terms of geometrical primitives (e.g. by the interpretation of RCF, the theory of real closed fields, into Tarski's geometry). The geometric formulation may simply advocate a purely geometric context theory (say, Tarski's planar axioms) and take the notions of area and size comparisons or sums of these to be geometrical primitives. Depending on the particular theory chosen, one would have to accept the fact that there are some definitional 'gaps' in the theory (see for instance (Beeson, 2023) for an analysis of the notion of 'equal figures', which is undefined in Euclid's theory).

Now the notion in the proof of Pythagoras's Theorem that should be decisive for the purity result is, of course, similarity. We defined similarity earlier as equality of the angle measures, or as equality of corresponding lengths of the sides of the triangles (using proportionality). In the arithmetical case, suppose we pick Tarski's geometry as our context theory, and define numbers using the interpretation of RCF. Then this interpretation will additionally be used to define the operations used in $A^2 = B^2 + C^2$, i.e. addition, and exponentiation (in terms of multiplication). Similarity can then easily be defined by taking the numerical lengths of the sides of the triangles, and by defining proportionality in terms of a definition of division (obtained by adapting the definition of multiplication: if $A(x, y, z)$ defines $x \cdot y = z$, then it also defines $\frac{z}{x} = y$). Thus, in this case, defining similarity draws upon the same ontological operations as the ones that are already needed to formalize $A^2 + B^2 = C^2$. In other words, the proof is restricted to the same ontology as that of the theorem, and it obtains *full ontological purity*.[9] Thus, this says that ontologically, similarity cannot be considered as distinct from an arithmetical reading of the theorem, in contrast to epistemic approaches to purity. Additionally, here secondary purity results can be established for any proof that 'rephrases' the similarity proof in terms of a different ontology, provided the theory capturing this other ontology can interpret the context theory.

Finally, suppose we pick the fully geometrical ontology. The proof then requires a geometrical theory of a certain strength to define the notion of similarity. Hilbert, for instance, establishes similarity using axioms I, 1-2 and II-IV of his planar theory (excluding the axiom of Archimedes). Generally, for the geometrical context theory that we pick, the question is then whether we think that the axioms necessary to define similarity are also necessary to properly describe the geometric plane that Pythagoras's Theorem talks about. This is not an obvious question, and the answers may differ per investigator. In this case, the safest choice for full ontological purity may be to pick a subset of a geometric theory that cannot

[9]This assumes that we think Tarski's geometry including its definition of numbers is considered to really refer to an ontology of the plane as well as arithmetic. One could of course pick other context theories: but full purity will still be achieved, as similarity becomes harmless once we already formalize the theorem in terms of arithmetical operations.

define similarity. This reflects the historical controversy surrounding similarity as presented in the literature, where when one thinks of the ontology of Pythagoras's Theorem, this does not include a way to know whether two triangles have the same angles or proportional sides. This says that even though a theory that defines similarity describes the right type of ontology for Pythagoras's Theorem (planar geometry), it does so with a complexity that is too high. The consequence of this is that the proof is *not* fully ontologically pure — instead, as its counterpart, it will become *fully ontologically impure*. Additionally, the proof will not obtain any *secondary ontological purity*. This is because secondary ontological purity requires that any notion in the proof is first of all formalizable in the context theory (so that in a theory interpreting the context theory, the proof can be restricted to just simulations of the pure ontology), while in this case, similarity is not.

Depending on the arithmetical or geometrical reading of the theorem, then, we get a purity or an impurity result. This shows that similarity (and proportionality) more naturally belongs to an ontology that includes arithmetic, while it is more keenly avoided in a purely geometric ontology.

## 2.5 Models of explanation

We will here discuss the theoretical ingredients of three models of explanation:

Kitcher (1989)'s *explanation by unification* (an epistemic model)

Steiner (1978a)'s *explanation by a characteristic property* (an ontic model)

Pincock (2015)'s *explanation by abstraction* (an ontic model)

These models provide an individual take on what the relation between a mathematical theorem and its explanatory proof amounts to. After describing their approaches in more detail, we discuss how they fare on the pure proof of the Infinitude of Primes.

#### Kitcher's model of explanation

Among the several existing contemporary models of explanatory proofs, one of the most well-known is Kitcher (1989)'s account of explanation, which originates from the area of scientific explanation, and has been tested for mathematical explanation in (Mancosu and Hafner, 2008). Kitcher's model is a unificationist model of explanation, as it is based on the idea that explanations provide understanding by unifying different phenomena. His view is that "an explanation should make the best tradeoff between minimizing the number of patterns of derivation employed and maximizing the number of conclusions generated" (Kitcher, 1989). In what follows we summarize the way in which Kitcher makes these insights precise.

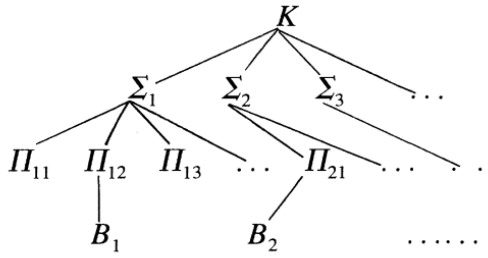



Figure 2.6: Visualization taken from (Kitcher, 1989)

We start by considering a fixed set of beliefs $K$, which can amount to a set of statements supported by a scientific community, or more specifically to mathematics: a set of theorems in a certain mathematical theory. $K$ is assumed to be consistent and deductively closed. The beliefs within $K$ can then be deductively *systematized* by a set of arguments, i.e. by derivations of sentences in $K$ from other sentences in $K$. The systematization of $K$ that is most unifying will be considered most explanatory, and will be given by the *explanatory store* over $K$, $E(K)$.

We describe how to find the explanatory store similarly to (Mancosu and Hafner, 2008). First, we define three notions. A *schematic argument* is a sequence of schematic sentences, namely sentences that require some non-logical expressions to be replaced by dummy letters. Second, a set of *filling instructions* will specify exactly how these dummy letters can be substituted again by expressions with content. Third, a *classification* for a schematic argument tells us which sentences in the argument are premises, which ones are conclusions, and by which rules conclusions are inferred from premises. Then a *general argument pattern* is a triple $(s, f, c)$ consisting of a sequence of schematic sentences $s$, a set of filling instructions $f$, and a classification $c$ for $s$.

We still need to introduce a few more notions to be able to rank argument patterns according to their unifying power. A set of derivations is *acceptable* relative to $K$ if every step is deductively valid, and each premise belongs to $K$. If $\Sigma$ is some set of derivations (a *systematization* of $K$), then a *generating set* for $\Sigma$ is a set of argument patterns $\Pi$ such that each element of $\Sigma$ instantiates some pattern in $\Pi$. Now a generating set $\Pi$ for $\Sigma$ is *complete* with respect to $K$ if every derivation that is acceptable relative to $K$, and which instantiates a pattern in $\Pi$, belongs to $\Sigma$: so within our context $K$, all derivations exemplifying a pattern in $\Pi$ need to be contained in $\Sigma$ (see Figure 2.6 for a visualization of the relation between these sets).

Now we can determine how unifying such a systematization is, and find the most unifying one among different acceptable systematizations, $E(K)$. First, for each acceptable systematization, take all the corresponding generating sets $\Pi$ that

are complete with respect to $K$. Out of these generating sets, subsequently, select a *basis*: the generating set that contains the least number of patterns. Finally, rank all of the bases according to their unifying power. For this, we should also define the *conclusion set* of $\Sigma$ ($C(\Sigma)$, containing the sentences that occur as the conclusion of some argument in $\Sigma$). Then the degree of unification of a systematization $\Sigma$ is describable by the ratio $\frac{|C(\Sigma)|}{|\text{basis}|}$ (although Kitcher himself provides a more qualitative judgement of the level of unification), and $E(K)$ becomes the systematization where this ratio is the highest.

**2.5.1 Example.** A simple example from mathematics is given by Mancosu and Hafner (2008). Consider the problem of determining the equation of the line tangent to the parabola $y = 2x^2 + 3x + 1$ at point $(1, 6)$. We can solve the problem using derivatives as follows.

1. $2x^2 + 3x + 1]' = 4x + 3$
2. $[4x + 3]_{x=1} = 7$
3. Thus the tangent line to $2x^2 + 3x + 1$ at $(1, 6)$ is $(x - 1)7 = (y - 6)$.

A schematic argument for determining the tangent line to a differentiable curve $f(x)$ at point $(x_0, y_0)$ can be obtained from the above as follows.

$1S$. $[f(x)]' = g(x)$

$2S$. $[g(x)]_{x=x_0} = c$

$3S$. Thus the tangent line to $f(x)$ at $(x_0, y_0)$ is $(x - x_0)c = (y - y_0)$.

We can then define appropriate filling instructions by specifying how to replace $f(x), g(x)$ and $c$, and clarify that $1S$ and $2S$ are premises, while $3S$ follows from the premises by calculus.

**Steiner's model of explanation**

An established ontic model of mathematical explanation is that of Steiner (1978a). Steiner starts from the idea that we look for the 'essence' or 'nature' of an entity, when we try to explain its behaviour. In the context of mathematics, he makes the notion of 'essence' precise by what he calls 'characterizing properties'. A *characterizing property* is "a property unique to a given entity or structure within a family or domain of such entities or structures", where the notion of 'family' is taken as undefined.

An explanatory proof should then satisfy two properties. First, it should refer to a characterizing property of an 'entity' mentioned in the theorem. This characterizing property should be essential to the proof of the theorem: "[i]t must be evident, that is, that if we substitute in the proof a different object of the same

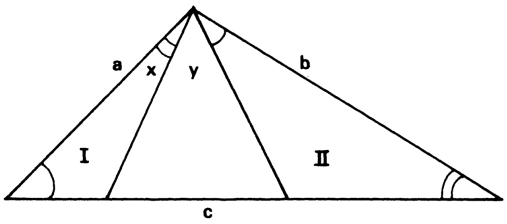



Figure 2.7: Visualization taken from (Steiner, 1978a)

domain, the theorem collapses". This ensures that the theorem really depends on the characterizing property.

Second, any explanatory proof should be *generalizable*. This means the following to Steiner: if we systematically alter the specifics of the characterizing property in such a proof, this should give rise to an 'array' of corresponding theorems. That is, creating such 'deformations' of the original proof should give rise to corresponding variants of the theorem. This means that explanation becomes a relation between a multitude of proofs and theorems.

**2.5.2 Example.** Consider the proof of Pythagoras's Theorem in Section 2.2.1. The proof uses a characterizing property of right triangles: they are triangles that are decomposable into two triangles similar to each other and to the entire triangle. Other triangles do not satisfy this property. The theorem can additionally be seen to depend on this property (see Figure 2.7): "[i]f we let the vertex of the right triangle vary (calling the largest side of the triangle, $c$, the hypotenuse), and try to decompose the triangle as before, by drawing lines $x$ and $y$ from the vertex to $c$, such that triangles $I$ and $II$ remain similar to each other and to the whole, we find that triangles $I$ and $II$ fail to exhaust the whole when the vertex varies between 90° and 180°; overlap when the vertex diminishes from 90° to 60°". The theorem then does not hold anymore: instead, (Steiner, 1978b, p.137) shows that it then becomes $c^2 = a^2 + b^2 - 2ab \cos C$ ($C$ being the angle opposite side $c$). This immediately gives us a generalization of Pythagoras's Theorem, where any 'deformed' triangle gives a precise instantiation of the latter equation. Right-angled triangles then turn out as the unique case where $2ab \cos C$ equals zero, i.e. where lines $x$ and $y$ coincide.

### Pincock's model of explanation

The last model of (ontic) explanation that we discuss is relatively more recent, and is described in (Pincock, 2015). Pincock's model aims to account for those explanations that provide a theorem's grounds. The most important ingredient of the approach is the existence of biconditionals, which "link facts of type $X$ to

facts of type $Y$" (Pincock, 2015). Types are intended to describe objects of a certain mathematical domain, whose properties can be described by different 'facts'. Facts of type $X$ with constituent $x$ are denoted by '$X_i(x)$', where the index $i$ distinguishes different facts of the same type $X$. Pincock's notion of explanation then first requires there to be a biconditional between facts of different types that occur in the proof. This biconditional only holds when objects of the different types are paired up in a certain way by a relation $R$. More concretely, the first requirement is that for any two objects $x, y$, whenever the relation $R(x, y)$ holds, the biconditional between facts $X_i(x) \leftrightarrow Y_j(y)$ also holds. The idea is that this biconditional reflects the "objective dependence relation" (i.e. the grounding relation backing the explanation) between facts of type $X$ and facts of type $Y$.

In order for the biconditional to fully represent the grounding relation, a second condition needs to be satisfied. For this, Pincock introduces a notion of abstractness: the constituents of the facts of one type must be *more abstract* than the constituents of the facts of the other type. The more abstract objects can then be seen to form the grounds for the less abstract objects. By an ordering of abstraction, Pincock means the idea that "some objects can have other objects as instances" (the instance being less abstract than what it instantiates), which is closely related to the type-token distinction.

In particular, furthermore, "the explanatory grounds will be given by facts of type $X$ whose constituents $x$ are more abstract than $y$ but less abstract than objects $z$ drawn from any other potential explanatory grounds": they are the *least* more abstract domain where the biconditional occurs. Pincock admits that demonstrating the 'leastness' of the abstraction difference is a challenging aspect of his account of explanation, and there is no reliable way of knowing this for sure.

**2.5.3 Example.** Pincock illustrates his approach by a proof of the 'unsolvability of the quintic': polynomials of degree $5$ are not solvable by radicals. That is, for equations $a_n t^n + ... + a_1 t + a_0 = 0$, for $n = 2, 3, 4$, $t$ corresponds to a formula outputting the values (*roots*) for which the equation is true. 'Solvability by radicals' means that this formula involves only the domain of the coefficients and the operations of addition, subtraction, multiplication, division and $\sqrt[b]{a}$.

The fact that for polynomials of degree $5$ this cannot be done, has a proof from Galois theory. In short (for more details, see (Pincock, 2015)), the proof uses the notion of a *field* and *field automorphisms*, and the fact that for each field extension, there is a collection of field automorphisms that forms a *group*. The step to polynomial equations is made as follows: polynomials are solvable by radicals just in case a radical extension of the field $\mathbb{Q}$ exists that includes the roots. This, in turn, can be determined by using the group of field automorphisms. In particular, the Galois group is the group given by the field extension by five roots. It can be shown that the field extension whose Galois group is $S_5$ is *not* a radical extension, so that these fifth-degree equations cannot be solved by radicals.

This fits into Pincock's model of explanation as follows. The dependence rela-

tion concerns facts about groups and facts about polynomial equations. Although polynomials are not instances of groups, each polynomial equation gives rise to a field extension and so a collection of automorphisms. And each collection of automorphisms *is* an instance of a group. Here, to be a group just means to be an abstract structure satisfying the group axioms. Instead, the collection of field automorphisms has additional properties concerning their being automorphisms of that particular field extension. The biconditional 'for any $x, y$, given $R(x, y)$, $(X_i(x) \leftrightarrow Y_j(y))$ is made as follows: whenever $x$ is the Galois group of polynomial equation $y$, $x$ is a solvable group if and only if $y$ is solvable by radicals.

### 2.5.1 Application to the pure proof of the Infinitude of Primes

This subsection will discuss the following results.

> The pure proof of the Infinitude of Primes is *explanatory* according to Kitcher (1989)'s unification explanation.
>
> The pure proof of the Infinitude of Primes is *unexplanatory* according to Steiner (1978a)'s dependence explanation.
>
> The pure proof of the Infinitude of Primes is *unexplanatory* according to Pincock (2015)'s abstraction explanation.

#### Kitcher's model of explanation

In order to test the pure proof of the Infinitude of Primes against Kitcher's model of explanatory proofs, we will define our set of beliefs $K$ to be the set of all arithmetical truths (as approximated by, e.g., the theorems of Peano Arithmetic (PA)). Then $K$ contains the theorem IP, and any other PA-statement about the natural numbers. We should then look for an argument pattern used in Euclid's proof that can systematize $K$ in the best possible way, that is, we should look for $E(K)$.

A straightforward search for an argument pattern focuses on the induction schema, which is a main strategy used in the proof: induction intuitively unifies many results about the natural numbers. As mentioned before, we note that there is a debate in the literature on whether proofs by induction are explanatory (see for instance (Lange, 2009; Hafner and Mancosu, 2005; D'Alessandro, 2019)). Kitcher himself belongs to those who believe induction can be explanatory (where below induction is used to show all numbers have 'property $F$'):

> Suppose that I prove a theorem by induction [...] we feel that the structure of the positive integers is exhibited by showing how 1 has the property $F$ and how $F$ is inherited by successive positive integers; and, in uncovering this structure, the proof explains the theorem (Kitcher, 1975)

In our application of Kitcher's framework, induction will indeed come out as explanatory. Hence, although this result is still open to discussion within the mathematical and philosophical community, it is in any case in line with Kitcher's impression about the topic. We first propose the following argument pattern.

1. $\varphi(x)$ holds for $x = 1$.
2. If $\varphi(x)$ holds for $x = n$, then $\varphi(x)$ holds for $x = n + 1$.
3. $\varphi(x)$ holds for $x = n$ for each $n \in \mathbb{N}$.

Here, let the *filling instructions* say that $\varphi$ should be replaced by any PA-sentence.[10] For IP specifically, for instance, we would replace $\varphi(x)$ by the description "there exists a $b > x$ such that $b$ is prime". Moreover, the *classification* of the argument is as follows: 1 and 2 are premises, and 3 follows from 1 and 2 by applying the induction principle of PA.

Let $\Pi$ be the singleton set containing the induction argument pattern. Then it is a generating set for a set of derivations $\Sigma$, namely the set of all derivations leading to PA-theorems of the form 'for all $n$, $\varphi(n)$'.[11] Then $\Pi$ is certainly complete with respect to $K$: every derivation that is PA-acceptable, and that instantiates the induction pattern in $\Pi$, belongs to $\Sigma$. But how unifying is this systematization? Of course, there are many other ways of selecting a systematization of PA-derivations, and finding (complete) generating sets $\Pi$ of argument patterns for these. Of all these hypothetical sets $\Pi$, recall that we should compare their number of argument patterns to the size of their conclusion set. Without going into detail about other possible PA-generating sets, we note that the current one seems highly unifying. For one, our $\Pi$ is a singleton containing just one argument pattern; as it is already as minimal as possible, its basis will then remain this singleton. Furthermore, there is an infinitude of theorems in PA that can be proven by the induction argument pattern, and so that can be unified by this argument pattern. Hence, dividing $C(\Sigma)$ (infinity) by the size of the basis (one) will result in an infinitely high degree of unification of $\Sigma$. This alone should convince us that $\Sigma$ is a very reasonable candidate for $E(K)$. Hence, a proof that relies on the induction argument pattern should come out as explanatory according to Kitcher's model.

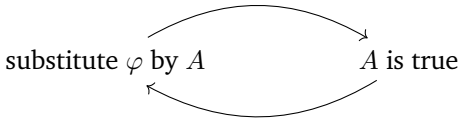
Let us dwell a bit more on some parts of our analysis. First of all, note that our filling instructions are quite loose, as they allow $\varphi$ to be substituted by any PA-sentence. At first sight, it may have been more intuitive to let the filling instruction be: "replace $\varphi$ by a PA-sentence for which each instance $\varphi(n)$ for $n \in \mathbb{N}$ is true (and which are provable by using induction)", as these are the sentences that we aim to have as our conclusions. However, choosing this more specific option

[10] The lack of constraints on the filling instructions is justified soon.

[11] $\Sigma$ contains all derivations that proceed by induction. Note that this equals all derivations that lead to theorems of the form 'for all $n$, $\varphi(n)$', since if such a theorem has a proof that does not proceed by induction, we can use this proof to form a trivial induction proof after all, so that this theorem is also generated by $\Pi$.

is problematic for at least two reasons. The first is that filling instructions do not stand alone but always come with a classification — and the classification, as noted by Mancosu and Hafner (2008), sometimes does the job of the filling instructions too. This is so in our case. It is the classification itself that does not allow the argument pattern to be instantiated by just any PA-sentence. It excludes the sentences for which the instantiation of 3 does not follow from 1 and 2 by induction. Thus, our filling instructions are allowed to be so general, because the classification overrides them. Then, opting for the more specific version of the filling instructions would in some sense be redundant.

Another reason to go for the more general filling instructions is to avoid a certain circularity in the argument pattern. If we use the more specific variant of the filling instructions mentioned above, we have to already know that $\varphi$ is a truth that holds for all natural numbers, *before* proving it. Kitcher is tolerant in what types of requirements can be put on the instantiations of dummy letters in argument patterns — but it seems that they should only allow restrictions on instantiations from $K$ that are epistemically *prior* to carrying out the proof. This is clearly violated in our example, where we have the following circularity:




This discussion about the level of generality of the filling instructions is linked to the two features of *stringency* and *spuriousness* that Kitcher puts forward as desirable and undesirable, respectively. Arguments instantiating a pattern are 'stringent' if they "contain some non-logical expressions and [...] are fairly similar in terms of logical structure". The classification ensures these similarities in logical structure, while the non-logical expressions in the argument pattern together with the filling instructions determine the way dummy letters are substituted. Presumably, both factors should secure a desirable type of argument that is suitable for unificatory explanation. But stringency can be 'mimicked' by imposing artificial restrictions on the substitution of non-logical expressions, while these are not actually essential — e.g. by using a very general argument pattern such as 'from $A$ infer $A$', and by letting the filling instructions substitute $A$ with whatever conclusion one desires. Thus, Kitcher tests the filling instructions on 'artificialness': if you can replace them by a new set of filling instructions, which then allows the derivation of *any* sentence whatsoever, the constraints on the instructions were actually artificial, and not really stringent. Such artificial argument patterns are then exactly what Kitcher calls *spurious* patterns, which should be avoided. Spurious unifications such as the one mentioned above unify a large number of results in a 'trivial' way, and should not be characterized as explanatory. Now as we saw before, spuriousness can be avoided by designing meaningful filling instructions, but also by making the classification more specific. As in our induction argument

pattern, the classification can override the filling instructions, and so prevent 'any sentence' from becoming derivable. Our argument pattern is thus not spurious, and seems Kitcher-explanatory.

We should remark that Kitcher also says: "I do not wish to claim that my requirement will debar all types of spurious unification". So, even though our argument pattern is not classified as spurious because it satisfies Kitcher's requirement on the filling instructions, there might still be some other desirable requirement (e.g., on the classification) to prevent spuriousness, one that our argument pattern does not satisfy. For instance, we might want to require that the classification of an argument pattern must focus solely on the logical structure of the argument, and that it somehow cannot contain 'too much' non-logical content. Whether this can be coherently specified is not obvious. Still, our result that Euclid's proof is explanatory with respect to the induction argument pattern, is compatible with Kitcher's own views on induction, and so seems a legitimate application of his framework. On a broader level, we may interpret the result of induction being Kitcher-explanatory as saying that explanation has many intuitive aspects, and that different models of explanation can focus on and single out particular ones. Kitcher's approach singles out the globally unifying value of explanatory proofs.

**Steiner's model of explanation**

We will consider the pure proof of the Infinitude of Primes and see whether we can isolate a characterizing property in that proof (and, if so, whether the proof is generalizable). Since the Infinitude of Primes mainly concerns natural numbers, two promising candidate characterizing properties in the proof arise: one based on induction, and one based on the Fundamental Theorem of Arithmetic. However, we will argue that they both fail to be characterizing properties in Steiner's sense. As we do not expect other suitable characterizing properties to exist, we take this as support for the idea that the proof is Steiner-unexplanatory.

**Induction.** Again, as the pure proof of the Infinitude of Primes proceeds by induction, we can consider this principle as giving rise to a characterizing property. However, Steiner himself (as opposed to Kitcher, interestingly) claims that induction cannot support explanatory proofs:

> "The proof by induction does not characterize anything mentioned in the theorem. Induction, it is true, characterizes the set of all natural numbers; but this set is not mentioned in the theorem." (Steiner, 1978a)

From this we can gather at least that Steiner does not *intend* his model of explanation to capture proofs by induction as explanatory. Possibly, this is motivated by the intuition that induction is a basic method of proof: it is a general approach

to statements about the natural numbers, that convinces us that each individual number satisfies some property, but does not seem to say much about the 'nature' or 'essence' of natural numbers themselves, which is what Steiner demands.

Despite Steiner's position, our analysis is not over as we, similarly to Hafner and Mancosu (2005), will take a more tolerant approach towards induction as a potential characterizing property. Hafner and Mancosu argue that, although Steiner wants to exclude induction from being a characterizing property, his model does not directly do so. Namely, suppose we maintain that induction cannot be a characterizing property because theorems never mention quantifier ranges (such as the set of natural numbers) and this mentioning is actually necessary. Then there are several properties that should inuitively be counted as characterizing properties (e.g., as described by Hafner and Mancosu, certain results of Kummer's test on the convergence of series), but are rejected by the model because they are not mentioned by the theorem. Hence, the need for explicit mentionings of sets of objects in theorems makes Steiner's approach *undergenerative*. Additionally, even if we do accept that the theorem (in this case, IP) does 'not mention' the *set* of natural numbers, one can easily rephrase it such that it does, and "it seems quite odd that Steiner's theory qua theory of the explanation of *proofs* should turn out to be so overly sensitive to what appears to be a rather minor detail in the exact wording of a theorem which doesn't affect its proof" (Hafner and Mancosu, 2005).

Hafner and Mancosu therefore set Steiner's position on quantifier ranges aside, and describe a positive application of Steiner's model to a proof by induction of the summation theorem. They show that induction here suffices as a characterizing property, and also that it is generalizable to create an array of new theorems. They conclude that the proof is explanatory according to Steiner's approach, and that this makes Steiner's approach *overgenerative*, as they still hold a strong intuition that inductive proofs should not be considered as explanatory.

This is not the place to decide whether Steiner's model should or should not include mentioning of quantifier ranges. Rather, we want to show that, even if we adopt Hafner and Mancosu (2005)'s approach to induction — i.e. even if we accept that the mentioning of quantifier ranges in theorems is not necessary to isolate characterizing properties — induction still does *not* come out as a characterizing property for the pure proof of IP. In other terms, unlike the proof of the summation theorem, the proof of IP turns out not to rely in Steiner's sense on the property of induction.[12]

Now consider why induction is not a characterizing property for the pure proof of the Infinitude of Primes. As in the proof, let $Q_n = p_1 \cdot ... \cdot p_m + 1$ for all primes $p_i < n$ $(i \in \{1, ..., m\})$. Then let the theorem of the Infinitude of Primes be phrased

[12]Incidentally, if induction *had* been a characterizing property for our proof of IP, it is interesting to note that it would also have been able to satisfy generalizability, in a similar way as the proof of the summation theorem in (Hafner and Mancosu, 2005). Indeed, we can change the notion of being 'prime' to variants 'prime*' that hold for different subsequences of the natural numbers, leading to new theorems. However, since induction is not even a characterizing property for us, this becomes irrelevant.

as in Section 2.2.1:

(IP) For each natural number $n$, there exists a prime $p$ such that $p > n$.

In order for induction to be a characterizing property, the proof must 'depend' on it. This means that "[i]t must be evident, that is, that if we substitute in the proof a different object of the same domain, the theorem collapses" (Steiner, 1978a). Like Hafner and Mancosu, we consider induction as potentially characterizing the set $\mathbb{N}$, and we take as a domain the family of sets in the powerset of $\mathbb{N}$ ($\mathcal{P}(\mathbb{N})$). Hence, we are investigating whether induction characterizes the natural numbers relative to this domain. As a candidate characterizing property, induction can then be phrased in terms of a free set variable $X$, so that the instantiation of $X$ by $\mathbb{N} \in \mathcal{P}(\mathbb{N})$ will certainly satisfy induction.[13]

But we see that dependence of the proof on the characterizing property is not satisfied: there exist several (infinite) subsets of $\mathcal{P}(\mathbb{N})$ and corresponding (restricted) induction principles that, when substituted in the proof, would lead to the same result. For instance, if we show that for all *even* numbers $n$, there exists a prime $p > n$ (by an induction principle with an induction step of $n + 2$, instead of $n+1$), then we may also conclude that for all $n$, there exists a prime $p > n$. The result does not depend on the individual numbers characterized by the induction principle: the principle just needs to characterize infinitely many of them. Hence, as the proof does not display the sort of dependence on induction that Steiner specifies, induction cannot be a characterizing property for the proof. Note that also considering a different type of family does not help, as any family of sets containing $\mathbb{N}$ arguably either contains an infinite subset of $\mathbb{N}$, or an infinite superset of $\mathbb{N}$ (such as the integers $\mathbb{Z}$). In any of these cases, we are still free to conclude the theorem IP.

To sum up, even when leaving open the possibility that induction might be a characterizing property for Steiner's account, we see that not all instances of induction can be seen as a characterizing property: it depends on the particular use of induction which occurs in the proof under scrutiny. In the case of the pure proof of the Infinitude of Primes, it turns out that induction is not a characterizing property.

**Fundamental Theorem of Arithmetic.** We briefly consider another candidate characterizing property for the pure proof of the Infinitude of Primes, which is based on the application of the Fundamental Theorem of Arithmetic (FTA). We can view FTA as characterizing the set $\mathbb{N}$, as we did for induction — but also as characterizing individual numbers $n$, as a unique product of primes (within, say, the set $\mathbb{N}$). Let us analyze this latter option first. If we go for it, we are bound to accept that the proof is dependent on the prime product of a particular natural

[13]The (slightly bulky) notation of induction in (Hafner and Mancosu, 2005) is given by the formula $1 \in X \& \forall x(x \in X \rightarrow (x+1) \in X) \& \forall P[(P(1) \& \forall x(P(x) \rightarrow P(x+1))) \rightarrow (\forall x \in X, P(x))]$.

number: i.e., if we replace it by another natural number with a different prime product, the proof should fail. But this is clearly not the case: the proof deals with an arbitrary natural number and its prime product. Thus, looking at FTA as characterizing individual numbers in the pure proof of the Infinitude of Prime is not a successful strategy.

Let us then consider FTA as characterizing the set $\mathbb{N}$, which we phrase in the following way (again with a free set variable $X$):

**[FTA]** Each $x \in X$ is either a unique product of prime numbers, or equal to $1$.

Now we can see quickly that we are not faced with a characterizing property within the family $\mathcal{P}(\mathbb{N})$. Indeed, if we replace $\mathbb{N}$ by a(n) (infinite) subset of $\mathbb{N}$, this subset still satisfies FTA, so FTA does not uniquely characterize just $\mathbb{N}$. The proof of IP could still use the FTA-property to obtain a higher prime as before, leading (for infinite subsets of $\mathbb{N}$) again to the result that there exist infinitely many primes. Thus, FTA also does not lead to a characterizing property in Steiner's sense. As there do not seem to be other suitable candidates for characterizing properties in the pure proof of the Infinitude of Primes, we take this as support for the idea that the proof is Steiner-unexplanatory.

### Pincock's model of explanation

Recall once more from Section 2.2.1 that Euclid's proof of IP proceeds by induction, and that, for the inductive case $n$, it constructs a number $Q$ that is the multiplication of all primes lower than $n$ plus $1$. The Fundamental Theorem of Arithmetic then gives us a prime that divides $Q$, and that is bigger than $n$. Given the form of the proof, at least at first sight no ground naturally emerges, and no intuitive abstraction difference seems to be at play. However, let us consider any biconditional that might arise in this setting. The domains (and facts about them) that the proof of IP arguably connects are the natural numbers and prime numbers. We are thus looking for a relation $R$ between natural numbers $x$ and primes $y$, such that if it obtains, then a biconditional connecting facts $X_i(x)$ about natural numbers to facts about primes $Y_j(y)$ holds.

In an attempt, let $R(x, y)$ say that $y$ divides $x$. We then claim that this forces the biconditional $X_i(x) \leftrightarrow Y_j(y)$ where $X_i(x)$ says that $x$ is a natural number, and $Y_j(y)$ says that $y$ is a factor of a prime product of $x$. To see that the biconditional works, consider both directions. From left to right, assume $y$ divides $x$ and that $x$ is a natural number.[14] Then we know that there is a number $n$ such that $y \cdot n = x$, and by FTA, either $n = 1$ so that $y^1$ is a prime product, or $n$ has a prime product $p(n)$, so that $y \cdot p(n)$ is still a prime product. (Note that the prime product of $x$ does not have to be unique here.) For the right-to-left direction, assume $y$ divides

[14]Recall that by definition of the domains $X$ and $Y$, $x$ must already be a natural number, so that the fact $X_i(x)$ is quite trivial — and that we also already know that $y$ comes from a domain of prime numbers.

$x$ and that $y$ is an element of a prime product of $x$. Here, $R(x, y)$ is not necessary to see that $x$ is a natural number: already by assuming that $x$ has a prime product, it must be the case that $x$ is a natural number (by definition of prime numbers). The role of $R(x, y)$ is thus vacuous for this direction of the biconditional.

Pincock intends the biconditional to form an objective dependence relation between $X_i(x)$ and $Y_j(y)$. But our attempt at a biconditional only conveys a trivial property: it says that if a prime divides a natural number, then it is contained in some prime product of the natural number. It would be reasonable to terminate the analysis here, as no significant biconditional can be extracted out of the proof. For the sake of exploring all aspects of Pincock's approach, we shortly explore any abstraction difference between the two domains linked by the biconditional. Here, we can make use of a more elaborate understanding of the notion of abstractness, as put forward by Marquis (2016) and which seems to further specify Pincock's perspective. In particular, according to Marquis's proposal, one of the main indicators of abstraction differences amounts to the fact that some properties that are used to distinguish the more concrete objects *disappear* at the more abstract level. This means that, when relating the two domains, multiple concrete objects should be identifiable with the same abstract object, so that the latter fails to distinguish with the same level of detail between concrete objects. Hence, when we move from the concrete objects to the abstract objects, some 'irrelevant details' of the concrete objects are forgotten at the abstract level; they are simply not expressible anymore.

Now, intuitively, there is no abstraction difference between the domains of prime numbers and natural numbers because, when moving from one domain to the other, we are merely 'cutting out' objects from the natural numbers, in order to be left only with the particular natural numbers that are prime. That would be like taking the domain of spheres, and cutting out ones of a certain size — or taking the domain of groups, and selecting from this the abelian groups. Marquis (2016) agrees that "the notion of group and the notion of abelian group [...] are just as abstract", because "[b]eing less abstract cannot merely be captured [...] by the fact that the extension of the [supposedly less abstract concept] is strictly included in the [supposedly more abstract concept]". But notice also that the requirement of properties 'disappearing' during abstraction is not satisfied: we are not forgetting any properties about prime numbers when we move to natural numbers. Admittedly, *more* numbers appear in the set of natural numbers that do not have the particular property of primality, but among the natural numbers it is still completely clear that the prime numbers *do* have this property. Finally, the type-token comparison, put forward by Pincock, also falls short: it is certainly not the case that multiple prime numbers are instances of one natural number. In short, on this analysis of abstraction, no biconditional linking prime numbers and natural numbers can truly satisfy Pincock's requirements.

Although the notion of abstraction is a complex concept, which surely admits different interpretations, by our findings that both the biconditional, as well as a

supposed abstraction difference between natural numbers and primes are controversial at best, we conclude that the pure proof of IP is *Pincock-unexplanatory*.

## 2.6 Reflections on purity and explanation

We have seen that purity and explanation, as characterized by a multitude of models, are ideals of proof with a pluralistic nature. Based on our findings, we will reflect some more on the models we applied, as well as on the supposed interaction between purity and explanation.

### 2.6.1 Models of ideals of proof

Models of ideals of informal proof are relatively abundant (for well-known ideals), but "[d]elicate judgments based on expertise may be involved in determining that something actually fits the metaphysical theory of [these ideals]" (Pincock, 2015). That is, there appears to be a gap between the theory of pointing out what metaphysical description can be given to an ideal of proof, and the practice of using the model to determine whether a given informal proof satisfies this ideal. Within this 'delicate' area of ambiguity, that in the end only practitioners of mathematics can bring true clarity to, we may comment on some findings regarding the relation of our selection of models to mathematical practice — in particular, these findings can be seen as interfering with a reliable relation. Such aspects are not only a nuisance: they also help exhibiting the nature of models of proof ideals, and of informal proof itself (and ways in which they lack the straitjacket of formal proofs, which we will concern ourselves with more in the rest of this dissertation).

First, as may be expected in the informal setting, the precise *wording* of a theorem, as well as of some ingredients of the models of proof ideals themselves, can affect the result of applying the model. This is a problem if this is unintended. In the case of our model of purity, the precise wording of the theorem affects the chosen ontology. However, as expanded on more in the next chapter, it is an intended feature of the model that the purity result is relative to the judgements of mathematicians relative to this choice. That is, the flexibility is built into the model. A similar situation is going on in topical purity, where the investigator has a lot of freedom in determining the commitments making up the topic of the theorem. This can have as a consequence that no decisive, independent purity result can be obtained by a neutral investigator. However, we still get a purity result *relative* to possible choices, that can subsequently still be made by the relevant investigator.

A different situation seems to concern 'unintended' reliance on precise wording, where the ambiguity obscures what is the 'correct' outcome of the model. The choice in interpretation here creates doubt around how the model is supposed to work: it is something that the model is supposed to determine, but unintentionally does not. Examples of this are arguably the notion of 'operation' in (Kahle and

Pulcini, 2017), that we have already seen in our discussion in Section 2.4.1 — but also the notion of an 'argument pattern' in (Kitcher, 1981), the notion of 'abstraction' in (Pincock, 2015). These models are not presented as taking the application result as relative to the expertise or intuition of the investigator. The theorem is supposed to 'mention' the relevant operations, argument patterns have a certain 'objective' existence, and abstraction is described by its properties independent of an investigator. Here, models of proof ideals reach their limits in practice. Additionally, consider (Steiner, 1978a), whose model also requires the characterizing property to concern entities that are *mentioned* by the theorem. The word 'mentioning' is highly susceptible to phrasing of the theorem, as shown by the case of induction. In general, then, there is simply a lot of freedom in applying any of these models to concrete mathematical proofs. This freedom can be incorporated into the model, but if this is not the case, then border cases become problematic.

Second, even if the workings of the model seem reliable and relatively well-understood, outcomes of the model can still be undesirable. That is, models of ideals of proof tend to be overgenerative, undergenerative, or both, based on examples that have intuitive clarity. While this is often seen as a weakness of models of proof ideals, we would here like to emphasize (as mentioned at several points earlier in this chapter) that models of proof ideals inevitably cannot be 'calibrated' completely right. For one thing, a model of an ideal often starts from an example proof in mathematical practice that provides an exemplary case of the ideal of proof — hence, the specifics of the model are likely to be biased towards an example of this kind. For another, a model simply has to make a specific choices representation of the ideal. Through the lens of a model, the proof ideal thus loses generality, and instead becomes a *distorted* representation of the original, intuitive ideal. We believe that this is inevitable for any of these models, but that this should not be interpreted as a bad thing. As claimed earlier, and in line with studies such as (Inglis and Mejía-Ramos, 2021), each model becomes an analysis of certain specific *aspects* of ideals of proof, and in doing so give body to different variants of such ideals. For instance, Kitcher provides a clear focus on the unificatory aspect of explanation. Although e.g. Pincock (2015) has noted that Kitcher cannot account for his example of the unsolvability of the quintic, we can interpret this to merely show that unification is not the *kind* of explanatory aspect that the unsolvability of the quintic example possesses. Rather, this example can be seen as possessing various other aspects of explanation, and it is insightful to see which ones, by developing models that can account for it (such as Pincock (2015)'s model).

In short, we recognize the shortcomings of models of proof ideals that follow from the informal level of analysis we find ourselves on. Simultaneously, we advocate pluralism in the sense that developing these models only provides more insight into the available precisifications of 'fluffy' ideals of proof, and how we can make sense of an array of variants of each ideal.

### 2.6.2 Theoretical interaction between purity and explanation

Now, taking our models at face value, we can consider their interaction, which we will first do based on the similarity between their theoretical ingredients. Out of the six models we selected, we might expect to see a pattern related to the division between epistemic and ontic models. That is, we might believe that epistemic models of purity and explanation provide similar judgments of proofs they are applied to, and similarly for ontic models. However, a closer theoretical observation suggests a more nuanced view.

First, the epistemic model of topical purity can be considered unrelated to Kitcher's epistemic model of explanation. Kitcher's model takes as its basis the similarity of the structure of a proof argument to as many other proofs as possible. This does not require any particular relation between the proof and the theorem itself, which is essential in the case of topical purity. In fact, such a relation between the proof and theorem seems required for any of our other models of purity or explanation, suggesting a theoretical separation of these models from the way that Kitcher's model characterizes explanation.

As for the comparison of topical purity to ontic models of explanation, no clear pattern arises. For one, Pincock (2015) proclaims that "abstract mathematical explanations are decidedly [topically] impure". This seems reasonable, as supported by Arana and Detlefsen (2011)'s classification of Furstenberg's proof of IP as impure, where sets and topological elements are taken as 'going beyond' the arithmetical context of IP. I.e., any mathematical object that does not contribute to an immediate understanding of the theorem, is considered as extraneous — while Pincock's explanation comes exactly from linking two entirely different mathematical domains. Finally, Arana (2022) analyzes the theoretical (and practical) relation between Arana and Detlefsen (2011)'s topical purity and Steiner (1978a)'s model of explanation. At first sight, it may seem that the notions of 'topic' and 'characterizing property' are quite similar. But Arana notes that the type of knowledge that these notions lead to is rather different. In particular, characterizing properties are properties of entities that are unique within a family of those entities. The property specifies how certain objects relate to other objects within this family. This is presented as a metaphysical truth, independent of an agent's epistemic status. A topic, on the other hand, consists of agent-selected commitments, that reflect an their semantic understanding of a theorem. In short, while these notions may happen to coincide, they certainly do not have to.

Consider now the theoretical ingredients of the ontic models, each of which connects a collection of certain mathematical objects or operations relating to the theorem, to those relating to the proof. In particular, note that Pincock and Steiner advocate different sources of explanation. For Pincock, explanation comes from linking each object relating to the theorem to a more abstract object, that has less properties. Steiner, on the other hand, starts with a family of entities, where the characterizing property (by selecting a more specific subset of this domain) does

explanatory work. The ontic models of purity relate in various ways to differences in specificity of objects. Operational purity allows the proof to contain other operations than the theorem, and distinct operations may be closed under different numerical domains. The differences in domains concern possible extensions of the same domain by new numbers. This suggests a potential (but superficial) similarity to Steiner's model: a domain restriction can do explanatory work for Steiner, while operations 'characterizing' restricted numerical domains can secure purity. There cannot be said to be any reliable relation, however, between characterizing properties and operational ontologies.

Furthermore, although our ontological modal of purity allows for arbitrary domain changes (for secondary purity), the level of abstraction or specificity with which objects are described remains the same, as any property of the original domain is translatable (able to be simulated) by the secondary domain. The idea is that the same proof can in theory be carried out by both domains. Possible similarities may be found with Pincock's model of explanation, that links each object in the domain of the theorem, to a more abstract object in the proof. However, this is intended to be a true abstraction relation, where objects in the theorem are instances of the more abstract ones occurring in the proof. In ontological purity, the interpretation translation induces just a one-to-one correspondence between objects, and it ensures that no properties disappear. That is, our notion of 'surrogate' is different from the notion of 'abstract entity' of Pincock. Surrogates still need a rather restricted, particular domain change — while Pincock's abstract objects can in principle possess any property. Thus, it is in the end unlikely that an explanation in terms of Pincock's model will possess any ontological purity.

All in all, we see several high-level, theoretical similarities between the models of explanation and purity, but when spelled out in detail, these do not seem to guarantee any similar outcomes in practice.

### 2.6.3 Practical interaction between purity and explanation

We end our reflections by diving some more into the interaction between purity and explanation, based on the findings from our case study. Consider again the results of our analysis as summed up by the following tables.

| *Models of explanation* | **Kitcher** | **Steiner** | **Pincock** |
|---|---|---|---|
| *Pure proof IP* | Explanatory | Unexplanatory | Unexplanatory |

| *Models of purity* | **Arana & Detlefsen** | **Kahle & Pulcini** | **Martinot** |
|---|---|---|---|
| *Explanatory proof PT* | Impure | Impure | Pure/impure |

As the tables show, more often than not, explanation and purity come apart. This confirms general impressions witnessed in works such as (Lange, 2015; Lehet, 2021), but also departs from the Aristotelian tradition mentioned in the introduction, that tends to see the explanatory power of a proof and its purity as two faces

of the same medal. As mentioned above, we can also partly attribute this general pattern to the development of new models in general, that zoom in on and 'distort' certain aspects of explanation and purity.

With a little more nuance, we specifically recognize several trends from the comparison of contemporary models. First, the purity of Euclid's proof of IP goes quite well together with Kitcher's *epistemic* model of explanation, but diverges from the *ontic* models of explanation. As suggested above, however, the ingredients of Kitcher's model i.e., the level of unification of a proof, independently of its theorem — are not necessarily relevant for purity decisions. We thus suggest to see the convergence of purity with Kitcher's approach mainly as a coincidence. Purity of Euclid's proof of IP is guaranteed by the theorem restricting itself appropriately to the arithmetical context of the proof. Kitcher's type of explanation is achieved because of the proof strategy of induction — these factors are independent of (but at least, still compatible with) each other.

Second, the divergence of purity with respect to ontic accounts of explanation can be read along the following lines. As mentioned already, Pincock's approach requires a more abstract domain to explain a less abstract domain, which conflicts with purity when seen as attempting to *avoid* abstraction differences (as these differences heighten the risk of extraneous elements to occur in a proof). Indeed, the pure proof of the Infinitude of Primes decidedly does not contain abstraction differences. For Steiner's model, we could not find a property characterizing the natural numbers such that the proof depended on it in Steiner's sense. This was a consequence of the generality of IP: other subsets of the family $\mathcal{P}(\mathbb{N})$ would also have contained infinitely many primes. Thus, domain changes involving Pincock's increase of abstraction, or Steiner's increase in specificity (by a characterizing property), are also relatively independent from purity considerations. Generally, then, similarity of objects and operations between a theorem and proof induces purity, while a domain difference is required for explanation.

Second, the explanatory power of the similarity proof of Pythagoras's Theorem diverges from an *epistemic* perspective, but also, at least partly, from an *ontic* perspective on purity. Recall that Steiner's explanation characterizes right triangles by the property of similarity, within a family of other triangles. Topical purity considers the notion of similarity already too complex, and irrelevant, for purity to hold. For operational purity, proportionality by division is the operation that outstrips the natural numbers, creating impurity. This shows that Steiner's notion of characterizing property thus disregards issues of (conceptual) complexity of its domain: a subset of the 'family of entities' may be selected by any property (as long as uniqueness and dependence are satisfied). As for ontological purity, the convergence between purity and explanation depends on the choice of ontology as geometrical as well as arithmetical (leading to purity), or as fully geometrical (leading to impurity). In the former case, the reason for purity is the easy formalization of the notion of similarity in such a theory. However, Steiner considers similarity to explain due to its uniqueness and the proof's dependence on this

property — again, matters of easy understanding or formalization are irrelevant here. So, in general, the reasons why a theorem is true in an explanatory proof are simply not guaranteed to restrict themselves either to a pure ontology, or to a pure epistemic context.

Finally, we come back briefly to the historical development mentioned in the introduction of this chapter, where we observed that purity and explanation know a close origin, but diverge more and more in modern approaches. Our study seems to be in line with this development. Generally, this can be explained by the simple fact of more detailed models being developed studying these ideals of proof, but may also come with the trend of mathematics where boundaries between disciplines of mathematics disappear. This raises the question whether there can still exist satisfying models of purity and explanation that do coincide in their core, emphasizing their similar traits. The recent account of mathematical explanation of Poggiolesi (2024) may suggest a notion that is more compatible with purity as in the Aristotelian tradition. The main idea in this account is that what characterizes mathematical explanatory proofs consists in an increase of conceptual complexity from their premises — which can thus be seen as the grounds — to their conclusions. The two main tasks of the model are (i) to identify those premises of the proof under scrutiny which could be seen as the grounds of the theorem, and (ii) to show that these premises are indeed less complex than the theorem itself. A complexity difference can manifest itself on different levels; for instance by relating elements in the proof by definitions, or by considering the set of objects that the elements denote, where elements are related if there is a *theorem* that establishes a connection between their sets of objects. For full details, we refer to (Poggiolesi, 2024).

## 2.7 Conclusion

In this chapter, we explored the connections between the explanatory power and purity of proofs. We applied the main models of both notions to an explanatory proof of Pythagoras's Theorem and a pure proof of IP, while distinguishing bottom-up from top-down approaches, and ontic from epistemic approaches. Practically all analyses emphasize the divergence of purity from explanation, as supported by a conceptual analysis of the compatibility of the theoretical model ingredients. Generally, we advocate a pluralistic understanding of ideals of proof, where each model selects specific aspects to focus on. We encourage future studies in philosophy of mathematics to analyze in different ways where connections between the various qualities of mathematical proofs lie.

# 3

# Formalizing an ideal of proof

## *Full ontological purity*

The work in this and in the following chapter may be considered a case study of the following, general question: *how do ideals of informal proofs transfer to the setting of formal proofs?* In order to investigate this, we zoom in on the ideal of purity. Our first aim in this chapter is to introduce a new conception of purity that was already touched upon in Chapter 2, called ontological purity. Our second aim is to characterize which (classical) first-order natural deduction proofs of a mathematical theorem satisfy the 'full' version of this type of purity. Subsequently, Chapter 4 will concern a similar two-faced approach towards an extension of full ontological purity, called secondary ontological purity. Hence, both full and secondary ontological purity will come with a criterion of purity for natural deduction proofs, which can be seen as contributing to addressing the gap between informal and formal proofs.

In particular, given an informal mathematical theorem, we will provide a way of determining its ontological content. Formal proofs that refer to the content of theorem will be called 'fully ontologically pure'. After a few short formal remarks in Section 3.1, Section 3.2 investigates a previously made connection between purity for formal proofs and cut elimination. Section 3.3 then describes the ontological content of a theorem. The way in which this type of content can subsequently be captured by a formal theory (the *'context' theory*) is given in Section 3.4. We then argue that definitional extensions of the context theory capture the same ontological content as the context theory in Section 3.5. This results in criteria for full ontological purity for formal and informal proofs, which are presented in Section 3.4.2 and 3.6. The work in this chapter corresponds to the first part of the publication (Martinot, 2024a).

## 3.1 Formal preliminaries

It suffices to make a few small remarks about formalities. We will use the standard classical first-order language, containing $\wedge$, $\vee$, $\rightarrow$, $\forall$, $\exists$, and $\bot$, where $\neg\varphi$ is defined as $\varphi \rightarrow \bot$. A signature is a set of predicate symbols and function symbols (where constants are nullary function symbols), and we use $\mathcal{L}$ to refer to the first-order language of a certain signature. First-order theories $\mathsf{T}$ are sets of axioms in a language $\mathcal{L}_\mathsf{T}$. Furthermore, $\varphi, \psi, \chi, ...$ range over formulas in a language $\mathcal{L}$, and $s, t, u, ...$ over terms. We write $\Gamma \vdash_\mathsf{T} \varphi$ to mean that $\varphi$ is derivable from assumptions $\Gamma$ and axioms of $\mathsf{T}$ in the standard first-order natural deduction calculus (see e.g. (Buss, 1998)). To clarify the language of a formula, we write $\varphi_\mathsf{T}$ to mean that $\varphi$ is a formula in language $\mathcal{L}_\mathsf{T}$.

## 3.2 Remarks on cut elimination

Purity is by origin characterized by the intuitions of mathematicians concerning *informal proofs*, by which we mean proofs as mathematicians conduct them in practice and present them in natural language interspersed with formal symbols (see Chapter 1). We saw the most well-known models of purity in Chapter 2, including the notion of *topical purity* of Arana and Detlefsen (2011) and *operational purity* of Kahle and Pulcini (2017). We here also mention Baldwin (2013), who describes a context-relative variant of topical purity, that checks whether, given a certain formalization and context of acceptable concepts, the proof introduces any notion outside this context by explicit definition. The latter is an example of a characterization of purity that concerns informal proofs, but that uses several formal concepts in determining whether the proofs are pure. A specific investigation into purity for formal proofs has to our knowledge only been conducted by Arana (2009), who concludes that a syntactic approach to purity is not desirable (although such an approach is also discouraged by others, e.g. Baldwin (2013) and Kahle and Pulcini (2017)).

We here discuss the investigation of Arana (2009) in a bit more detail. When considering purity criteria for formal proofs, one might be tempted at first to provide fully mechanical and syntactic criteria. These may tell us independently, given any formal proof, whether it is pure or not. However, we reinforce the view that mechanical conceptions of purity are unsuccessful, by taking as an example the case study of the subformula property by Arana (2009). Arana investigated the idea that cut elimination is a procedure for 'purifying' a formal proof. Here, given a formalized theorem $\varphi$, the idea is that a pure formal proof is one that restricts itself to the set of subformulas of $\varphi$, $\mathsf{Sub}(\varphi)$. Namely, the proof then quite literally only draws on that what is stated by the theorem itself.

There are some practical limitations to this characterization of pure formal proofs. For sequent calculi to which no 'proper' first-order mathematical axioms

are added, the procedure of cut elimination for a formal proof of $\varphi$ will guarantee that any formula in the proof is in fact a subformula of $\varphi$ (and hence, supposedly, that the proof is pure). Arana notes, however, that for any decent theory of mathematics, only 'free-cut elimination' is attainable. When a proof is 'free-cut free', some formulas may, instead of being a subformula of the theorem, be a subformula of one of the axioms used in the proof (see (Buss, 1998)). Hence, Arana rejects free-cut elimination for purity (and cut elimination for some variant incorporations of axioms in sequent calculi described by Negri and Von Plato (1998)), since the choice of axioms now comes into play — and the axioms are not guaranteed to capture the right content. We agree with the latter objection to this method of mechanical purity analysis, and we add another perspective to it. Namely, we do not only reject free-cut elimination because we do not know whether the axioms themselves are relevant to the theorem, but because we cannot be sure whether the axioms will be *used* in a pure way. This will become clear in Chapter 4, when we extend the formal notion of content.

Additionally, we point out that even if a proof is not just free–cut-free, but fully cut-free, cut elimination has several flaws as a measure for purity. First of all, the subformula property focuses on restricting the use of relation symbols and connectives of a formal language $\mathcal{L}$ in a proof. Its influence on constants and function symbols is weak: for $Q \in \{\forall, \exists\}$, $\mathsf{Sub}(Qx\varphi(x)) = \mathsf{Sub}(\varphi) \cup \{\varphi[t/x] \mid t \text{ a term of } \mathcal{L}\}$. This means that throughout the proof, in principle any constant or function symbol of $\mathcal{L}$ is allowed to instantiate a quantifier as part of a term. However, constants and function symbols typically formalize specific elements of the domain and operations (consider $0, +$ and $\cdot$ in Peano Arithmetic (PA)). The unrestricted use of constants and function symbols can therefore potentially lead to impurity in a proof, even if the latter satisfies the subformula property.

Secondly, as a consequence of the former point, the subformula property has a variable 'purifying' effect depending on the expressivity of a formal language. For a language that only has relation symbols, such as the signature $\{\in\}$ of ZFC, the subformula property is strong. No formula that has a different structure than the theorem with respect to the membership relation can occur in the proof — and ZFC has no choice but to describe every mathematical entity with this relation. On the other hand, the more expressive language $\{0, S, +, \cdot, <\}$ of PA has a weaker subformula property. As mentioned, it will not distinguish between instantiating $S0$, $0{+}S0$, $S0{\cdot}S0$, etcetera, in a proof, even though it the particular operations used may make a difference for purity considerations. Thus, the subformula property cannot be used as a universal purity measure for all formal theories.

Finally, we elaborate on a short point made by Kahle and Pulcini (2017) that the subformula property is quite strict and might exclude pure proofs. Even disregarding its weakness on constants and function symbols, the subformula property does not perfectly capture our conception of content. What the subformula property measures is strict syntactic similarity to a specific formal sentence. That does not guarantee a correspondence to (a possibly much more tolerant) intuition.

There are many different ways of formally describing a particular entity, and (with respect to relation symbols and connectives) the subformula property only allows very specific ones. For example, if the theorem is of the form $\varphi$, then the subformula property already prevents its derivation by $\wedge$-elimination from $\varphi \wedge \varphi$, even though conceptually $\varphi$ and $\varphi \wedge \varphi$ refer to the same mathematical objects and operations. So, although the strictness of the subformula property might guarantee purity for relational languages, it only captures a part of what formal proofs we think should be considered pure. Here, then, we aim for a more conceptually unifying approach to restricting the formulas in formal proofs.

Additionally, note that impure objects and operations can be described 'explicitly' (directly by a primitive from the language) or 'implicitly' (by a definable formula using lower-level primitives). If we prevent specific terms from occurring in a proof, we have not necessarily avoided the same notion occurring implicitly in the proof. Implicit definitions are hard to restrict: the subformula property does this by measuring strict similarity to the 'complexity' of relation symbols in the theorem, but as we saw this seems not fine-tuned to correspond to intuitive notions.

Instead, our approach will actively incorporate the intuitions of mathematicians into the context theory selection, and use formal tools only to more fully capture and refine these intuitions. We will now proceed with the introduction of ontological purity, where purity is generally achieved if any notion in a proof can be made sense of in terms of the mathematical ontology that a theorem concerns.

## 3.3 An ontological understanding of content

Generally stated, we interpret the content of a theorem as what the theorem is about. More specifically, we stick to the conception that a theorem is "about those things to which the terms appearing in it refer" (Detlefsen, 2008). This is one of several possible conceptions of purity, and puts the emphasis on mathematical material itself. We can think of content as the 'ontological realization' of the topic or subject matter of the theorem, i.e. the range of mathematical objects and operations that the theorem speaks about. By not yet introducing any axiomatic systems, syntax or semi-formal definitions, we aim to capture an intuitive view of mathematical material. For example, we may think informally about an ontology of 'the natural numbers', but not yet associate it with specific primitives or underlying principles. We think an intuitive way of considering mathematical material captures the essence of mathematical content, and this is what purity is naturally based on.

This view also ensures that (for now) we leave open the formal characterization of the content, emphasizing that a theorem is not ultimately about defining principles, but rather the things they define. On the other hand, the epistemological notion of purity of Arana and Detlefsen (2011) takes as a basis for content

(topic) "the elements that determine our grasp or understanding of mathematical problems", such as definitions, axioms and inferences. Such content is given to a certain problem $\mathcal{P}$, consisting of an interrogative attitude, a propositional content, and a formulation of the content. So, while their approach is also sensitive to different possibilities of formalization, Arana and Detlefsen accommodate this in an early stage — and additionally include particular formalization choices in the topic of the problem. For now (as far as possible) we will stay away from any specific formalization choices, and focus on the ontology itself.

The focus on ontology is also motivated by our intent to look at purity for formal proofs. How formal proofs correspond exactly to informal proofs, and how they can preserve epistemic values of informal proofs, is difficult to determine. It is far from evident that a proof system can be genuinely close to how mathematicians think in practice. This means that any truly epistemic notion of purity seems (for now) unsuitable for formal proofs. We think a more accepted premise is to let the syntax of formal proofs correspond to a mathematical ontology.

We see an ontology as a "domain of discourse", or a "realm of mathematical objects", as in (Shapiro, 1997). Given an informal theorem, its content is obtained by deciding what basic mathematical objects it is making a claim about. These should be objects in their intuitive form such as numbers, lines, classes, sets, and so on. They can also be particular variants of these sorts (even numbers, finite sets, etc).[1] For purity purposes, it is important to have a subjective feeling of the nature of these objects; and one should be able to describe the size of this domain (e.g., if a theorem talks about all numbers, the ontology should be infinite). For instance, the Infinitude of Primes informally stated as "for all natural numbers $a$, there exists a natural number $b > a$ such that $b$ is prime" concerns an ontology of all natural numbers. Additionally, the relevant complexity of these objects should be considered. What version of the chosen mathematical objects does the theorem concern, i.e., what main operational machinery are the objects equipped with? Arana and Detlefsen (2011) tell us that the content of the Infinitude of Primes is made up of axioms or definitions of successor, induction, an ordering, primality, and divisibility and multiplication, and that "the first-order Peano axioms for the natural numbers provide a reasonable formulation of these commitments, augmented by the definition of primality and divisibility". Namely, these ingredients can all be seen as necessary in order to properly understand IP. We agree that the intuitive operations behind these axioms or definitions are part of the ontology of IP, as they show the way the objects are related to each other and how they may be manipulated. Our notion subtly differs by leaving open definitional dependencies between operations, which fits the idea that an ontology can be described in multiple ways.

[1] An ontology can also be made up of abstract objects, such as groups or lattices, if they are considered as proper objects of a mathematical domain themselves, or as concrete instances of another mathematical sort: for instance, 'all finite sets that are groups'.

**3.3.1 Remark.** As remarked by a reviewer, we are essentially taking the notion of an (intended) standard model for ontological content. This is a helpful alternative description we have in mind, and it also aligns with the way Shapiro (1997) uses the term 'domain of discourse'. However, we will stick to the terminology 'ontology' throughout this paper, to emphasize that the intuition for an ontology precedes any formal theory, and that for any reasonably complicated theory it becomes problematic to choose a standard model. We will sometimes, however, use the 'standard model understanding' in places where a formal theory has already been chosen, and where it reinforces our argument, for instance in the next section.

## 3.4 A formal counterpart for ontological content

In order to let a formal proof correspond to ontological content, we take the notion of a formal (first-order) mathematical theory (as in Section 4.1) as the formal counterpart for content. Given an ontology, we will call this theory the *context theory* (see Figure 3.1). We recognize that, in principle, there is no necessary relation between a theory and a mathematical ontology in two ways: first, the subjective nature of the ontology of a theory is open (in an informal sense, PA can be taken to refer to numbers, but also to binary strings, sets, and so on). Secondly, a theory can have multiple potential ontologies (just like it can have different models), and an ontology can also be formalized by different theories, that prove different subsets of all the true sentences in the ontology. However, the decision to accept a relation between an ontology and a theory is what transfers the meaning of purity to the formal setting. Thus, we require such a relation to be fixed, in order to talk about purity for formal proofs. We think the incorporation of a theory choice is natural, as formal theories are commonly designed with the purpose of describing intuitive mathematical material. Consider for instance: "Geometry began with the informal ideas of lines, planes and points [...] Gradually, these were massaged into Euclidean geometry: a *mathematical theory* of these notions [...] [Peano Arithmetic] was intended to be a theory of our intuitive notion of number, including the basis of counting" (Turner, 2010). And purity judgements in practice already include drawing on sets of axioms, e.g. "[...] there are proofs, like Furstenberg's topological proof of the infinitude of primes, whose axioms are widely agreed to be irrelevant to the conclusion [...]" (Arana, 2009).

In Arana and Detlefsen (2011), notions like axiomatic theories, definitions, inference rules, and so on, mark certain epistemic commitments and make up content (a 'topic') itself, instead of providing a formal counterpart for content. Baldwin (2013) picks a formal vocabulary and theory to describe the topic, but only explicit definitions are compared to the intuitive content. For us, the context theory is itself a complete formal counterpart to an ontology, and we do not impose any restrictions on what the theory may define (note in particular that the theory

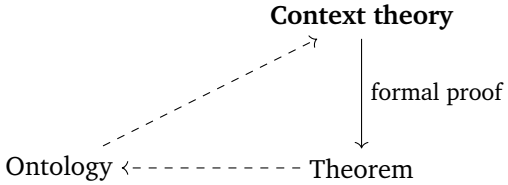



Figure 3.1: Visualization of the connection between the theorem and its (in)formal content.

also does not need to be able to prove the theorem, which allows for the prevention of purity results with respect to that theory).

Purity of a formal proof will then depend on whether the syntax in the proof indeed refers only to the right ontology. We think of a first-order theory as referring to an ontology through two aspects: the signature and the axioms. The signature of the context theory (constants, function symbols and relation symbols, and we here think of variables as well) denotes the basic objects and operations: terms will pick out objects, and function symbols and predicates operations and properties.[2] The referents of the primitives should correspond to basic elements and properties of the ontology, that intuitively determine its nature. However, objects and operations may also be referred to by descriptions in terms of these symbols, where their definition will determine their ontology. For example, in PA, the constant $0$ denotes the particular object we think of as 'zero', and $S$ an intuitive successor relation. Then $S0$ will not denote an isolated, separate object that we think of as 'one', but it will really correspond to 'one' as equivalent to 'the successor of zero'. Hence, the ontology of the primitives will determine the ontology of more complex syntax. Furthermore, after Quine, quantifiers will signal ontology by indicating reference to an object in the domain: "[a]n object exists, or is in our ontology, just in case it is in the range of a bound variable" (Shapiro, 1997). What kind of intuitive objects exactly make up this ontology, however, is left up to the interpreter of the signature of a theory.

Secondly, the axioms of a theory play a role in referring to the ontology. Theories with the same signature can have different ontologies, simply because their axioms differ. For example, depending on the axioms, set theories with just the signature $\{\in\}$ can have the cumulative hierarchy $\mathsf{T}$ as their ontology, or $\mathsf{T}$ including urelements, $\mathsf{T}$ plus an inaccessible cardinal, and so on. If a theorem speaks about all sets, the ontology in this case depends on what the interpreter considers to be the 'right' universe of sets. If a theory is considered to refer to a proper part of the ontology of a theorem, however, it can refer to nothing extraneous, and it

[2]The reference to ontology can thus be seen as similar to model-theoretic interpretation of a signature — while emphasizing that predicates, function symbols and terms should be seen as referring to the 'actual', intuitive objects, operations and properties we have in mind, instead of just tuples of domain elements that satisfy them.

will also preserve purity of proof. That is, although one can maintain one collection of mathematical entities as the 'real' ontology a theorem speaks about, purity of proof should be preserved for restrictions of that ontology.

The exact way that theories refer to mathematical objects and operations is a field of research of its own, with various problems described, for instance, by Shapiro (1997); Lavine (2000), such as whether a formal theory always has enough syntactic labels for ontologies with very large domains. Again, we will not assume any necessary ontological commitments of theories, however, but we ask some relation between an ontology and a theory to be fixed. Whether this relation is reasonable can be verified by the mathematical community.

### 3.4.1 More on the selection of a context theory

The previous section has generally clarified how to select an ontology for a theorem, and how a theory can subsequently refer to this ontology. We here highlight two aspects of choosing a context theory, that give some more insight into how the method works in practice.

First, the selection of a context theory can be dependent on who you ask: while most of us may find a theorem that concerns the natural numbers to correspond to an arithmetical theory like PA, a set theorist may really think of the numbers as sets, and connect the ontology of natural numbers immediately to a set theory restricted to the domain of set-theoretic natural numbers. Similarly, mathematicians may pick theories with different strengths. To illustrate, take the simple theorem "there are no two consecutive even numbers", where the notion of being even can involve division by two or being the sum of two equal numbers. One person may prefer the first conception, and pick PA as context theory, while someone who prefers the second conception may as well pick the weaker theory PrA (Presburger Arithmetic), which does not define multiplication. Finally, we may only have weak intuitions about the ontology of some theorems. For example, a simple set-theoretic theorem as Cantor's Theorem (for all sets $A$, $|\mathcal{P}(A)| > |A|$) concerns all sets, but does not clearly tell us which universe of sets (described by which axioms) it is about. Rather, this seems dependent on what one considers the main set-theoretical universe (this is mirrored by the unclearness on what the standard model of set theory should be). These situations all show that individual preferences play a role, and that the resulting notion of purity should be seen as relative to the individual choosing the context theory.

Second, a theory capturing a certain ontology may be able to encode other content. This relates to what Isaacson (1987) calls 'hidden' content of a theory, which may represent potentially extraneous elements, such as in the example below.

**3.4.1 Example (Convergence of two Taylor series).** Described in (Lange, 2019) is the theorem that the Taylor series of $1/(1-x^2)$ and $1/(1+x^2)$ have the same convergence behaviour. Lange notes that the mathematical objects that the theo-

rem refers to are Taylor series that involve real numbers. In particular, complex numbers can be considered impure elements, while their introduction provides a natural and explanatory proof. Any theory of the real numbers (such as Tarski's first-order axiomatization of real closed fields, RCF[3]), will be able to represent complex numbers as ordered pairs of real numbers.

We maintain that such a coding of extraneous elements does in fact not reduce the suitability of a context theory for capturing an ontology. Namely, we fix a relation between an ontology and the context theory (say, RCF). When RCF represents complex numbers, we have a choice in how to interpret these representations ontologically: as pairs of real numbers, or as a separate ontology of complex numbers. The fixed relation of RCF to an ontology of real numbers justifies that we interpret them as real numbers, as we have not assumed any connection between complex numbers and RCF. Thus, it is characteristic of the ontological approach that if we consider the ontology of real numbers acceptable, we find pairs of real numbers acceptable, too (possibly in contrast to an epistemic approach to purity). In Section 3.5, we repeat this point by making an extension of context theories. Complex numbers can then only be considered impure if they are thought to not 'really' be pairs of real numbers, but to be separate objects of their own. In this case, we may still recognize that the pairs of real numbers can 'accurately approximate' complex numbers. In Section 4.2.1, we will say more about such approximations, and we will attribute a secondary level of purity to them.

### 3.4.2 First criteria for full ontological purity of proof

We now present a first criterion for full ontological purity of formal as well as informal proofs. The criterion for formal proofs will be extended in Section 3.6. We do not claim any relation between an informal proof and a particular formal one, and so the purity results for formal and informal proofs should be seen as separate, though of course, compatible.

Given an informal theorem, the previous sections can be seen to provide the components of what we refer to as the *ontological context* of that theorem, denoted as a tuple $(O, \varphi_{\mathsf{T}}, R)$. In this context, $O$ indicates the choice of ontology for the theorem and $\mathsf{T}$ the choice of context theory, where $\varphi_{\mathsf{T}}$ stands for the theorem formalized in $\mathcal{L}_{\mathsf{T}}$. We introduce $R$ as a specification (with a reasonable level of detail) of how the signature of $\mathsf{T}$ naturally captures the basic elements of the ontology, as elaborated on in Section 2.2.

For formal proofs, we restrict to the standard first-order natural deduction calculus. We need to be convinced that any inference rule that we use cannot introduce concepts outside of our ontology. As we consider the inference rules of

[3]A reviewer notes that it is more accurate to say that RCF refers to the real algebraic numbers, or even another ontology — and that RCF cannot prove all intuitive properties of the real numbers. RCF should thus be seen as an imperfect context theory choice, but still as one of the most suitable first-order theories we have to refer to the 'real numbers'.

the first-order natural deduction system to be truly logical, we consider them as satisfying this requirement. This then justifies the following criterion for full ontological purity.

**3.4.2 Definition (First criterion for full ontological purity of formal proofs).** Given an informal mathematical theorem corresponding to the ontological context $(O, \varphi_{\mathsf{T}}, R)$, any formal proof $\Gamma \vdash_{\mathsf{T}} \varphi_{\mathsf{T}}$ is *fully ontologically pure* for that theorem.

For proofs as actually carried out by mathematicians, we need a slightly different approach. Like formal proofs, an informal proof should be fully ontologically pure if it only draws on the ontology of the theorem. This requires a relation between the notions described in an informal proof, and their interpretation in an ontology. Given an ontological context $(O, \varphi_{\mathsf{T}}, R)$ for a theorem, we may judge this by checking how the notions in an informal proof are formalizable into $\mathsf{T}$. What exactly is formalization is a question outside the scope of this paper, but we suffice here in saying that it cannot just be any mapping from informal notions to syntactic elements — it will have to satisfy certain criteria that convince us it is really a particular notion that is being formalized. Next, we assume that there is a difference between a formalization in general, and a *'natural formalization'*. We will say that a 'natural' formalization of an informal notion into $\mathsf{T}$ requires the notion to intuitively be made up of basic elements of the ontology of the signature of $\mathsf{T}$ (just like the connection of an ontology to a formal theory in Section 3.4)[4], and to syntactic descriptions in $\mathcal{L}_{\mathsf{T}}$ that are relatively efficient and/or elegant. Thus, we propose the following definition for full ontological purity, and discuss an example below.

**3.4.3 Definition (Criterion for full ontological purity of informal proofs).** Given an informal mathematical theorem corresponding to the ontological context of $(O, \varphi_{\mathsf{T}}, R)$, an informal proof of the theorem is *fully ontologically pure* if there exists a natural formalization of any notion in the proof into $\mathsf{T}$.

**3.4.4 Example (Infinitude of Primes).** Consider Euclid's proof (as described e.g. in (Arana and Detlefsen, 2011)) and the topological proof of IP (Furstenberg, 1955). Let the ontology of IP be the natural numbers with arithmetical operations including addition and multiplication, and let PA be the context theory. Any notion in Euclid's proof is formalizable in PA, and it should be relatively uncontroversial to say that any notion is even naturally formalizable into PA. We therefore consider this proof to be pure. The topological proof contains some elements that are by definition *not* formalizable in PA: for instance, the proof includes a topology of arithmetical sequences that has uncountably many elements. This is something that PA cannot define. However, it is still something that PA can *represent*, by

[4]We thus assume that besides associating a theorem to an ontology, one can intuitively associate notions in a proof to elements of an ontology.

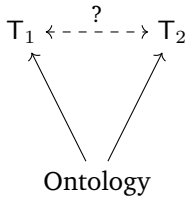



Figure 3.2: Equivalence of context theories.

letting an individual element stand for the uncountable set. We consider it likely for mathematicians to agree that this representation is still a formalization of the topological proof in PA. Although the representation of an infinity is simplified, the result is still recognizable as the notion occurring in Furstenberg's proof.

Whether the proof is fully ontologically pure, however, depends on whether the formalization in PA is natural enough. For a natural formalization, notions like 'arithmetic sequence' and 'integer' should intuitively be made sense of in terms of natural numbers, and correspond naturally to primitives or formulas of PA. In PA, an arithmetic sequence is represented by the coding of one of its elements $a + bn$. An integer $a$ or $b$ is represented for instance by an even natural number if it is negative, and an odd natural number if it is positive. But it should be clear that integers, or arithmetic sequences are not intuitively made sense of this way, in terms of the ontology of natural numbers. This is enough to conclude that Furstenberg's proof is not fully ontologically pure.

This shows that full ontological purity behaves relatively similar to traditional notions of purity. If the proof contains notions that are only naturally made sense of in an ontology that the context theory does not capture, this prevents full ontological purity.

## 3.5 Equivalence of context theories

The rest of this chapter is devoted to extending the first criterion of full ontological purity for formal proofs. We would like to consider formal proofs modulo differences that do not affect their capturing of content. Thus, we are looking for a notion of equivalence for context theories (see Figure 3.2). We will argue that definitional extensions provide such a notion. First, we introduce definitional extensions as in (Hodges, 1993).

**3.5.1 Definition (Explicit definition).** Let $\mathcal{L}$ and $\mathcal{L}^+$ be languages with $\mathcal{L} \subseteq \mathcal{L}^+$, and let $R$ be a relation symbol, $c$ a constant and $f$ a function symbol of $\mathcal{L}^+$. Then *explicit definitions of $R$, $c$, and $f$ in terms of $\mathcal{L}$*, respectively, are sentences of the form

$$\forall \overline{x}(R(\overline{x}) \leftrightarrow \varphi(\overline{x}))$$
$$\forall y(c = y \leftrightarrow \psi(y))$$
$$\forall \overline{x}, y(f(\overline{x}) = y \leftrightarrow \chi(\overline{x}, y))$$

where $\varphi, \psi$ and $\chi$ are formulas of $\mathcal{L}$.

**3.5.2 Definition (Definitional extension).** Let $\mathsf{T}$ be a theory of language $\mathcal{L}$. A *definitional extension* of $\mathsf{T}$ to $\mathcal{L}^+$ is a theory $\mathsf{T} \cup \{\theta_S \mid S \text{ a symbol in } \mathcal{L}^+ \backslash \mathcal{L}\}$ where for each symbol $S$ in $\mathcal{L}^+ \backslash \mathcal{L}$:

- $\theta_S$ is an explicit definition of $S$ in terms of $\mathcal{L}$.
- If $S$ is a constant defined by $\psi$ then $\vdash_{\mathsf{T}} \exists_{=1} y \psi(y)$. If $S$ is a function symbol defined by $\chi$, then $\vdash_{\mathsf{T}} \forall \overline{x} \exists_{=1} y \chi(\overline{x}, y)$.

Definitional extensions thus allow us to define abbreviations in a theory, by "replacing complicated formulas by simple ones" (Hodges, 1993). For example, in set theory we may denote the set $z$ such that $\forall w(w \in z \leftrightarrow w \in x \vee w \in y)$ by $x \cup y$. We will say that two theories are equivalent, if they are both *definitional extensions of the context theory*.

### 3.5.1 Referring to the same content

We claim that formal proofs from a definitional extension $\mathsf{V}$ of the context theory $\mathsf{T}$ are fully ontologically pure, because $\mathsf{V}$ refers to the same ontology as $\mathsf{T}$. Given the ontological context $(O, \varphi_{\mathsf{T}}, R)$, this is easiest to see if we interpret $O$ as the (intended) standard model of $\mathsf{T}$. The explicit definitions of an added symbol $S$ in $\mathsf{V}$ tell us exactly how to interpret $S$ model-theoretically in $O$: if $S$ is a constant, the explicit definition of $S$ will point to an object already in $O$. If $S$ is a predicate or function symbol, its explicit definition will point to relations between objects in $O$ that $\mathsf{T}$ was already aware of. Indeed, nothing new is added to $O$, only some elements of $O$ that were already there are given a new name.

One might raise the objection that adding an abbreviation can change the content: for example, suppose we extend PA by the definition for a membership-like symbol $\in$ from $\mathsf{ZF}^+_{\text{fin}}$ (see (Kaye and Wong, 2007)), or suppose we extend RCF by a relation symbol and definition for 'being a complex number'. In these cases, it seems we only introduced the abbreviation in order to talk about actual set-theoretic membership or actual complex numbers — and this appears to introduce new objects and properties we did not have before. For both (Arana and Detlefsen, 2011) and (Baldwin, 2013), it is the case that, for instance, adding an explicit definition of 'membership' to PA can introduce impurity in a proof of, say, IP. Something new is introduced that goes beyond the topic of a theorem, a concept that we ourselves can only really understand as set-theoretic membership.

However, for ontological purity of formal proofs, we emphasize that an ontological context $(O, \varphi_{\mathsf{T}}, R)$ gives an ontological interpretation of anything that $\mathsf{T}$

can prove. This contrasts with an epistemic conception of purity, where we may think the axioms of T belong to the content of a theorem, but it does not follow that we can easily understand (and thus accept) anything that T can prove. Ontological purity of a formal proof tells us that any notion in the formal proof has an interpretation made up of the basic ontological elements of the T-primitives. Any definitional extension certainly satisfies this, as its added symbols can be made sense of ontologically exactly the way T could make sense of their definitions. It is this interpretation of the added symbols that we claim is ontologically pure, and so this relies strongly on the connection of introduced symbols to their explicit definitions. Thus, when we add $\in$ to PA as in (Kaye and Wong, 2007), we are not 'really' adding set-theoretic membership to PA: we are abbreviating a complex number-theoretic property of PA, and we are simply affirming that this property is pure. The fact that this property simultaneously is a way of representing the membership symbol of $\mathsf{ZF}^{+}_{\text{fin}}$, does not take away the purity of the property in PA. The symbol $\in$ will only stand for actual set-theoretic membership, then, when it has a definition in, or is incorporated in the axioms of, a set theory that we associate with an ontology of sets. On the contrary, a definitional extension of PA by $\in$ gives an arithmetic definition of membership; and the syntax of this definition is made sense of by the PA-axioms, which were associated to an ontology of natural numbers.

This point is reinforced by the form of natural deduction proofs in a definitional extension. The formal proofs of the context theory are all preserved by the definitional extension, but we gain ones that use a definitional axiom as in Definition 3.5.1. A proper use of a definitional axiom can only serve to introduce the new symbol $S$ in the proof, which is initially embedded in a universal operator and a bi-implication. In order to actually use $S$ in a proof, first a proof of its definitional formula is required, and this will be given from the context theory axioms. This emphasizes that the definitional axiom is not a fully independently functioning axiom, and the extension cannot fashion the (ontological) meaning of $S$ out of thin air. This ensures purity of formal proofs from definitional extensions of the context theory.

### 3.5.2 Natural formalizations into definitional extensions

The situation for informal proofs again deserves a separate elaboration. We claimed that informal proofs that are naturally formalizable in the context theory, are fully ontologically pure. We here claim further that a proof is naturally formalizable in the context theory just in case it is naturally formalizable in a definitional extension of the context theory. Thus, while the inclusion of definitional extensions renders a larger number of formal proofs fully ontologically pure, it does not change which informal proofs are fully ontologically pure.

We illustrate our argument by taking again the example of the Infinitude of Primes. Let PA be the context theory, and consider the definitional extension

PA + $\forall x \forall a \forall b(E(x,a,b) \leftrightarrow \varphi(x,a,b))$. Here, let $E(x,a,b)$ be the relation symbol added to the signature of PA, that codes $x \in B_{a,b}$ for an arithmetic sequence $B_{a,b}$ as occurring in the topological proof of IP.[5] Intuitively, it feels like the notion of arithmetic sequences is more naturally formalizable in the definitional extension than in PA: it can now be efficiently formalized as a single predicate. At first sight, the topological proof of IP thus appears to become more pure with respect to the definitional extension.

However, note again that the predicate $E$ is tied to its explicit definition in the language of PA. That is, when we consider how to formalize $x \in B_{a,b}$ in the definitional extension, we cannot 'just pick' $E(x,a,b)$. We only know to pick this symbol as the formalization because it stands for the right arithmetical coding. The syntax of this coding, subsequently, is made sense of through their use in the axioms of PA, which were associated with an ontology of natural numbers. Thus, in the end it is always the fundamentally primitive syntax which determines a formalization and an ontology. The primitives in $\varphi(x,a,b)$ still refer naturally to arithmetic notions, and not to topological concepts. In other words, we still cannot say that Furstenberg's proof is fully ontologically pure in the definitional extension. So, we make the extension just to capture more broadly which formal proofs are fully ontologically pure.[6]

### 3.5.3 Other notions of equivalence

There exist many different notions of equivalence for theories, which can all be seen to potentially affect the nature of purity in a different way. Notions like synonymy, mutual or bi-interpretability (Friedman and Visser, 2014), which typically arose to transfer formal results between fields, are some other options. However, these notions would equate PA and $\mathsf{ZF}^{+}_{\text{fin}}$, and many more theories that we intuitively think of as capturing different ontologies. Another possibility is requiring that a theory is a conservative extension of the context theory. This choice, too, is more controversial: for instance, NBG, a set theory that includes classes in its ontology, is a conservative extension of ZF. Hence, for our notion of purity, we stick to the weaker notion of definitional extensions.[7]

[5] Here, $a$ and $b$ should also be seen as codes, for integers.

[6] This deserves emphasis to avoid misunderstanding: while formal proofs in definitional extensions of the context theory are pure because of their referral to a pure ontology, this does *not* mean that informal proofs using the notion that is intuitively abbreviated in the definitional extension, also become pure, as the intuitive notion may naturally concern a different ontology.

[7] Of course, one may still think that definitional extensions of a theory intuitively capture a different ontology as well — but we argued in this section that it is reasonable to abandon that intuition, given an ontological conception of content. For stronger notions of equivalence, we think this is no longer reasonable.

## 3.6 Extended criterion for full ontological purity of formal proofs

We end the chapter by presenting the extended criterion for full ontological purity of proof. The requirement for informal proofs here will remain the same as in Section 3.4.2, so the extension only has an effect on purity for formal proofs. We restrict again to the first-order natural deduction calculus, and first present a definition.

**3.6.1 Definition (Definitionally equivalent formulas).** Let $\mathsf{V}$ be a definitional extension of $\mathsf{T}$, extended by the symbol $S$ defined by the $\mathsf{T}$-formula $\psi$.[8] Then $\varphi_{\mathsf{V}}$ is *definitionally equivalent* to $\varphi_{\mathsf{T}}$ if:

- $\varphi_{\mathsf{T}}$ does not contain any instances of $\psi$, and $\varphi_{\mathsf{V}} = \varphi_{\mathsf{T}}$.
- $\varphi_{\mathsf{T}}$ does contain instances of $\psi$, which are possibly replaced by the abbreviation $S$. That is, we have the following cases:[9]
  - $S$ is a relation symbol with explicit definition $\forall \overline{x}(S(\overline{x}) \leftrightarrow \psi(\overline{x}))$, and $\varphi_{\mathsf{V}} = \varphi_{\mathsf{T}}[\psi(\overline{x})\backslash S(\overline{x})]$.
  - $S$ is a function symbol with explicit definition $\forall \overline{x}, y(S(\overline{x}) = y \leftrightarrow \psi(\overline{x}, y))$, and $\varphi_{\mathsf{V}} = \varphi_{\mathsf{T}}[\psi(\overline{x}, y)\backslash (S(\overline{x}) = y)]$.
  - $S$ is a constant with explicit definition $\forall y(S = y \leftrightarrow \psi(y))$, and $\varphi_{\mathsf{V}} = \varphi_{\mathsf{T}}[\psi(y)\backslash (S = y)]$.

Now for the criterion for full ontological purity of formal proofs.

**3.6.2 Definition (Criterion for full ontological purity for formal proofs).** Suppose we are given an informal mathematical theorem corresponding to the ontological context $(O, \varphi_{\mathsf{T}}, R)$. Let

$$[\mathsf{T}] := \{\mathsf{V} \mid \mathsf{V} \text{ definitionally extends } \mathsf{T}\}$$

Then for $\varphi_{\mathsf{V}}$ definitionally equivalent to $\varphi_{\mathsf{T}}$, any formal proof $\Gamma \vdash_{\mathsf{V}} \varphi_{\mathsf{V}}$ for $\mathsf{V} \in [\mathsf{T}]$ is *fully ontologically pure* for that theorem.

## 3.7 Conclusion

In this chapter, we have supplied formal and informal proofs with a notion of full ontological purity based on ontological content. For formal proofs, we suggest

[8] This definition can easily be extended to work for extensions by multiple symbols.

[9] We describe here the case where all instances of $\psi$ are replaced by $S$, but alternatively, $\varphi_{\mathsf{V}}$ may also replace only some or no instances of $\psi$.

that the context theory guarantees full ontological purity. Definitional extensions preserve this type of purity, because the symbols they introduce have definitions in terms of the primitives of the context theory, and so in terms of ontological elements that the context theory refers to. For informal proofs, full ontological purity requires 'natural' formalizability in the context theory. We argued that informal proofs that are naturally formalizable in the context theory, are automatically also naturally formalizable in its definitional extensions, and the other way around — hence, only one criterion for full ontological purity was required for informal proofs. These are consequences of the ontological approach we take towards purity (instead of, for instance, an epistemological one).

The framework for full ontological purity is flexible with respect to several aspects, the details of which a given investigator of purity can decide for herself: given a theorem, she can make the preferred choice of ontology; given an ontology, she can pick the preferred context theory; given the context theory, she can pick a preferred notion of equivalence for this theory; and finally, she can also interpret what it means for an informal proof to be 'naturally formalizable' into a formal theory. Hence, although we proposed specific interpretations of these concepts in this chapter, they remain open to different preferences. By leaving room for individual intuitions in the analysis of purity for formal proofs, we retain the connection to the informal nature of purity as an existing ideal. This way, we provide a way to bridge the gap from informal to formal proofs, without falling prey to inaccuracies of fully syntactic measures. The next chapter will extend the work done so far to a secondary level of ontological purity.

# 4

# Formalizing an ideal of proof

## *Secondary ontological purity*

As a continuation of the previous chapter, we will here introduce secondary ontological purity as an extension of full ontological purity, and elaborately describe which natural deduction proofs satisfy this type of purity. This effort again addresses the way that the ideal of purity is brought to the setting of formal proofs, but with the emphasis on extending the criteria for full ontological purity, so that we obtain a broader and more tolerant characterization of ontological purity. Simultaneously, we may view this effort as a case study for how proof-theoretic criteria can help identify philosophically meaningful derivations.

In particular, we will introduce the notions of surrogate ontological content and structural content of a given mathematical theorem. Formal proofs that refer to a surrogate version of the ontological content of a theorem will be called 'secondarily ontologically pure', because they preserve the structural content of a theorem. We present some formalities in Section 4.1. First, then, ontological content is extended to surrogate ontological content (see Section 4.2.1) and structural content (see Section 4.2.2). Surrogate and structural ontological content broaden the notion of traditional purity and blur some distinctions that are made in purity evaluations in practice, most clearly distinctions between disciplines of mathematics. Such a conception of purity has been mentioned before in the literature (see, e.g., (Arana and Detlefsen, 2011)), but has not before been made precise. We propose that this is a worthwhile extension in Section 4.2.3. Section 4.3.1 introduces interpretations between theories as in (Visser, 1997), and we describe how an interpretation can give rise to a restriction on syntax that allows for reference to surrogate and structural content in Section 4.3.2. We characterize the initial formulas and inference rule applications of natural deduction proofs that

satisfy this syntax restriction in Section 4.3.3, leading finally to the criterion for secondary ontological purity of formal proofs in Section 4.4.1. Secondary ontological purity for informal proofs is considered in Section 4.4.2. Finally, we comment on the interaction between full and secondary ontological purity in Section 4.4.3, and give (partial) criteria for impurity of proof in Section 4.4.4. The work in this chapter corresponds to the second part of the publication (Martinot, 2024a).

## 4.1 Formal preliminaries

As several sections of this paper will draw on rigorous definitions, we describe a few technical conventions here. As in Chapter 3, we will use the standard classical first-order language, and the derivability relation of the standard first-order natural deduction calculus.

Additionally, a derivation in the natural deduction proof system of theorem $\varphi$ is a tree $(V, E)$ labelled with formulas. Given a natural deduction proof and its tree representation $(V, E)$, we say that any node $v$ in the tree is *instantiated* by the corresponding formula in the natural deduction proof. The root is instantiated by theorem, and the formulas instantiating the leaves are axioms, or assumptions that are later discarded. Two nodes $v_{n+1}$ and $v_n$ are connected by an edge $(v_{n+1}Ev_n)$ just in case the instantiation of $v_n$ is obtained in the proof from the application of a single inference rule to a set of premises including the instantiation of $v_{n+1}$. Then a *branch* of the proof is any sequence $v_nEv_{n-1}E...Ev_1Ev_0$, where $v_n$ is a leaf, and $v_0$ the root. A branch is *open* when its leaf is instantiated by an assumption (that is either later discarded or not), while a branch is *closed* when its leaf is instantiated by an axiom.

Finally, we will also refer to a natural deduction proof by $\mathcal{D}$. Similarly to Visser (1997), given a formula $\delta$ with one free variable, $\delta_\varphi$ will stand for $\bigwedge\{\delta(x) \mid x$ free in $\varphi\}$. We then use $\lambda_\mathcal{D}$ to stand for $\bigwedge\{\varphi \mid \varphi$ instantiates a node in the tree representation of a proof $\mathcal{D}\}$. Thus, $\delta_{\lambda_\mathcal{D}}$ will be the conjunction of $\delta$'s applied to all free variables in a proof $\mathcal{D}$. This is not to be confused with our additional notation of $\delta_{\overline{x}}$. We will write $\overline{x}$ for the finite sequence of elements $x_1, ..., x_n$. Then $\delta_{\overline{x}}$ will stand for the conjunction $\delta(x_1)\wedge...\wedge\delta(x_n)$.[1] Thus, we maintain a difference between using a formula or a sequence of terms as a subscript. We will add smaller formal specifications in the relevant sections where necessary.

## 4.2 Extending ontological content

In Chapter 3, we have claimed that the mathematical ontology of a theorem can be captured by a theory and its definitional extensions, and that natural deduction proofs starting from the axioms of these theories (and informal proofs naturally

[1] We use this notation for conciseness, especially for the Appendix, Section 4.6.

formalizable in them) are fully ontologically pure. We are now interested in extending this notion of purity in a way that has been mentioned in the literature several times. For example, Arana (2009) mentions in a footnote that we might think that for Furstenberg's proof of IP, "the allegedly extraneous topological elements are really just reconceptualizations of what was already the concern, namely sets of natural numbers, and hence are in fact relevant to the infinitude of primes". Arana and Detlefsen (2011) note that Colin McLarty, as well as a Bourbakiste tradition of arithmetical research, consider theorems like IP to contain *both* arithmetical and topological content — although Arana and Detlefsen do not follow this line of thought, because on their epistemological account, a notion that takes into account the understanding of the theorem has priority. Additionally, McCarthy (2021) discusses the possibility that "[w]hether a proof of a number-theoretic assertion counts as "pure" depends on the conceptual background against which it is formulated. If it is framed in a wide context, for example, second-order analysis or ZF, then it may be that the most natural notion of purity is that applying to the wider context and not the strictly number-theoretic one".

Generally, these proposals are rejected by their authors, because they fail to retain the traditional values of purity. Hence, it seems clear that if we make such a generalization of purity, the original values of (im)pure proofs change. One may then have concerns about the motivation for considering this notion of purity. In the next section, we develop the notions of surrogate and structural content, after which we provide several reasons why the type of purity that deals with these forms of content is still worth studying. We will then introduce the notion of interpretations, which will serve to preserve the structural content of a theorem. This will enable a proof from a theory that is neither the context theory nor a definitional extension, when restricted by a derivation criterion (defined in Section 4.4.1), to still correspond to the structural content of a theorem and gain a secondary level of purity. We end the chapter by considering impurity of proof and the interaction between full and secondary ontological purity.

Here, we extend the notion of ontological content to surrogate content, which will be the ontology that secondarily ontologically pure proofs refer to. Next, we suggest the ontology of a context theory and its surrogate versions have structural content in common. This is what will justify the attribution of a secondary level of purity to proofs referring to surrogate content.

### 4.2.1 Surrogate content

Suppose that we relate a particular mathematical ontology to a context theory. Subsequently, we consider a second theory that concerns a very different ontology. Intuitively, we may ignore a large part of the latter informal objects and forget some of their properties, i.e., we may conceptually 'trim' and weaken the entities that make up the content for us. That is, (1) we can ignore the part of the domain

of the ontology of a second theory, that will not take part in representing the ontology of the context theory. We may (2) pair up or equate remaining objects if that is necessary for representing individual context theory objects. And (3) we can ignore properties of the objects that the second theory can prove, but that do not take part in representing properties of the ontology of the context theory.[2]

Arguably, we change the nature of the ontology of the second theory by carrying out these steps, since they do not form anymore the intuitive material that corresponds to the full formal theory. In fact, the ontology of the second theory can be (informally) restricted in such a way that the things that remain function as *surrogates* of the intended mathematical entities of the context theory. This makes up a large part, for example, of the foundational role of set theory:

> (Maddy, 2019) "To say that 'the universe of sets is the ontology of mathematics' amounts to claiming that the axioms of set theory imply the existence of (surrogates for) all the entities of classical mathematics – a simple affirmation of set theory's role as Generous Arena."

For instance, if we consider a universe of sets to be the material that ZFC refers to, we can restrict this universe to just the finite ordinals, that satisfy only set-theoretically realized arithmetical properties. The resulting intuitive objects can be seen as surrogates for the intuitive natural numbers that PA refers to — and we say that the restricted content of ZFC consists of surrogates of the content of PA. We emphasize that the extracted surrogate content in such theories really loses some of its original nature. That is, the collection of sets that simulate numbers is not by itself (without a connection to the full content) anymore the content of ZFC. Besides foundational theories, examples of 'surrogative reasoning' (a term also used in Swoyer (1991)) can be found in mathematics in practice: for instance in analytic geometry, where Cartesian coordinates are used to represent points in space. Additionally, the relevance of restricting to surrogates can be recognized in Maddy (2019)'s description of Essential Guidance: "[the universe of sets] includes hordes of useless structures and [...] no way of telling the mathematically promising structures from the rest".

In a footnote (pp.3–4), Arana (2009) suggests 'reconceptualizations' (similar to the notion of surrogates) may not be relevant for purity: first, "not every reconceptualization of a subject matter is necessarily relevant to that subject matter". Specifically, "the infinitude of primes does not seem to concern sets at all; so I do not see why sets, even simple ones, are relevant to the problem". We agree that surrogate sets are not necessarily relevant epistemically to IP: they do not seem to contribute to the usual understanding of the problem. However, we propose a different kind of relevance: surrogates are relevant for ontological purity, because they are connected through an underlying structure with the ontology of the context theory (see Section 4.2.2, and again, the motivation for such a notion

[2] On the view of an ontology as an intended standard model, surrogate content can be given a more precise description. We will comment on that in Section 4.3.1.

is elaborated on in Section 4.2.3). Another point made in the footnote is that "if Furstenberg's proof were formalized in a different way, say in a theory in which the notion of open set was taken as primitive [...] the response [that set theory shows the topology used to prove IP is just limited to simple sets] would not work. Why should set-theoretic formalization take precedence over these alternatives?" We think the emphasis should not be so much on the idea that the topology is limited to simple sets, but rather on the idea that the topology is limited to some representation that we know relates to arithmetic (such as 'simple', or surrogate, sets). In a theory that takes open sets as primitive, we could again find that the open sets used to prove IP are just simple ones, by seeing that they are still surrogates of arithmetical notions. That is, the aim is not to get rid of topology, but to make sure its use is restricted to number-surrogates, whichever ones.

The type of purity we will attribute to proofs referring to surrogate content has a 'secondary level' compared to full ontological purity. Namely, unlike definitional extensions of the context theory, for theories referring to surrogate content we cannot talk about strict equality to original content anymore, as they may well correspond to a different ontology. Thus, full ontological purity cannot be attained here. Secondary ontological purity then entails that, if a theorem concerns an ontology of natural numbers, its proof should only draw on this ontology, or alternatively its set-theoretic surrogates, geometric surrogates, and so on — but not anything else.

### 4.2.2 Structural content

We here propose that for a certain ontology, each type of surrogate content has structural content with it in common. That is, we suggest each theorem has structural content in the form of a mathematical structure, which has as its instances the ontological content of a theorem, as well as surrogate versions of the ontological content. Structural content has been proposed before as "the instantiation of a particular fundamental mathematical structure by the entities intuitively involved in the statement" Ryan (2021), but for us the structural content will refer exactly to the structure itself, where "[a] structure is the abstract form of a system, highlighting the interrelationships among the objects, and ignoring any features of them that do not affect how they relate to other objects in the system" (Shapiro, 1997).

There are several variants of structuralism, with a distinction between eliminative structuralism (there are possible structures, but not actual ones) and non-eliminative structuralism (there are actual structures) (Shapiro, 1997). Within non-eliminative structuralism, we can distinguish between in re structuralism and ante rem structuralism. In short, in re structuralism says that there is no more to structures than their instances, while ante rem structuralism claims that structures exist independently of the systems that realize them. Our approach to purity is neutral with respect to the debate on structuralism, but will adopt ante rem

structuralism for the notion of structural content. This is because it shows that, by existing independently from (non-mathematical) exemplifications, structures are themselves something that can be preserved when switching between ontological content and its surrogates. A structure independently shows the properties that different versions of surrogate content have in common with each other.

More specifically, "an ante rem structure is, or is akin to, an ante rem universal, in that it is a one-over-many. The same structure can be exemplified in multiple systems, and the structure exists independent of any exemplifications it may have in the non-mathematical realm. The difference between structures and the more usual kind of universal, such as properties, is that structures are the forms, not of individual objects, but of systems, collections of objects organized with certain relations" (Shapiro, 2008). Shapiro calls the structures studied in mathematics "free-standing", i.e. anything at all can occupy their places. Here, we are interested in all instantiations of the places and relations of ante rem structures by mathematical ontologies. Given the ontology of a context theory, its underlying ante rem structure can be seen as consisting of 'abstract places' for each object in the domain, connected to each other by 'relations' that determine the fundamental connections of the structure. Like Shapiro (2008), we think that "understanding the (formal) languages of mathematics is sufficient to understand the places and relations of at least some structures". For instance, the familiar natural number-structure consists of an initial object, and a successor relation satisfying the induction principle, connecting the initial object to a successor object, and so on.

A(n) (surrogate) ontology can let their objects occupy the structural places, and the structural relations can be occupied by their concrete instantiations selecting objects from an ontology. By allowing instantiation from any ontology, the structural content underlies ontological content. Its preservation when moving between surrogate versions of content is what justifies a secondary level of purity: it ensures that, while a proof can draw on various ontologies and disciplines of mathematics, it is at least restrained to the particular structure that a theorem concerns.

### 4.2.3 The value of extending ontological purity

We now have a conception of the types of content we want to extend purity results to. We here provide some reasons for why this extension of purity is worth considering. Arana and Detlefsen (2011) mention the *intervenient* value of traditional purity, as well as an *epistemic* value that they consider in more detail. The intervenient value of purity is broadly the "development of a thorough way of thinking" (Bolzano, cited in Arana and Detlefsen (2011)). We believe secondary ontological purity still encourages a variant of this benefit: it encourages one to, within a specific discipline of mathematics, focus ones thinking on specific surrogates only. The epistemic value, however, arguably disappears: this value focuses on giving

insights that have a certain simplicity and naturalness, and according to topical purity, reduces 'specific ignorance'. The latter is something our extension does not preserve, as we will for instance allow the concept of set to be used in a proof of IP, whereas Arana and Detlefsen (2011) do not, on account of its failure to reduce specific ignorance for IP. Then what other values can we attribute to our extension of purity?

First of all, the extension can be seen as the weakening to a 'core' notion of purity, one that tells us what pure proofs should satisfy at the very least. In other words, secondary ontological purity in a sense underlies other, more specific notions of purity that satisfy additional criteria — and perhaps this may help us understand the connections and dependencies between other types of purity better, by seeing them as particular cases of the extension. We also suggest that this makes it easier to distinguish *levels* of (im)purity, instead of the more blunt distinction between 'pure' and 'not pure'. We suggest we have secondary ontological purity and full ontological purity, but stricter notions like Arana and Detlefsen (2011)'s topical purity induce even stronger notions of purity, allowing for a more nuanced picture suggesting purity really consists of a family of notions.

Furthermore, the extension of ontological purity caters to the views of structuralists. For ante rem structuralists, traditional purity may not properly reflect what a theorem is about. Mathematical content may for them ultimately concern structures, and we suggest there is still a notion of purity for them to value. Similarly, the extension of purity also allows for the view that an informal theorem does not have one 'true' ontology, and that surrogate content is a relevant subject matter of a theorem.

We additionally claim that the extension of purity will still have an epistemic value. Traditionally impure proofs have been said to lead to new insights, by connecting mathematical notions that at first seem separate (see e.g. (Lehet, 2021)). Our extension of purity brings nuance to this value: it suggests that a pure proof can also have this value, but only by connecting notions from different mathematical ontologies that represent the same placeholder in a structure. On the other hand, impure proofs can still lead to new insights, but in a different way: by showing what intricate notions really go beyond representing 'pure' content, yet are still useful for proving a theorem. That is, the new distinction between purity and impurity allows us to see what parts of a theory can be seen as 'purer' than others.

Finally, it should be observed that the extension of purity is valuable for characterizing purity for formal proofs: separate from the motivation for introducing surrogate content, the setting of formal proofs itself already encourages an extension. For example, if we insist on saying something about the formal proofs of PA, but the context theory for our theorem is Presburger Arithmetic (PrA), then it is reasonable to think that a restricted version of PA can provide pure proofs of the theorem. Intuitively, after all, both PA and PrA refer to the natural numbers, only PrA assigns fewer properties to its numbers than PA. If we can exclude the

properties that PA can prove and that PrA cannot, proofs of PA can only draw upon relevant notions. Formal proofs can only satisfy such a restriction if we extend what we have done so far. We thus consider the extension to surrogate (and structural) content a reasonable one.

## 4.3 Formalizing extended content

In what follows, we introduce a way to use interpretations between theories to restrict formal proofs, so that they refer only to surrogate (and structural) content and thereby induce a secondary sense of purity.

### 4.3.1 Interpretations

Interpretations between first-order theories provide a way for theories to represent each other's language and provable statements. They will be important in developing formal guarantees for proofs to refer to surrogate content. We take a slightly altered version of the definition of relative interpretations in (Visser, 1997). Consider first-order theories $\mathsf{T}_1$ and $\mathsf{T}_2$ with respective languages $\mathcal{L}_{\mathsf{T}_1}$ and $\mathcal{L}_{\mathsf{T}_2}$ that have identity. An *interpretation* $i$ of $\mathsf{T}_1$ into $\mathsf{T}_2$ ($i : \mathsf{T}_1 \to \mathsf{T}_2$) has two ingredients, which will determine the interpretation translation $(.)^i$:

1. A function $F$ mapping the relation symbols $R$ and the function symbols $f$ of $\mathcal{L}_{\mathsf{T}_1}$ on formulas of $\mathcal{L}_{\mathsf{T}_2}$. If the arity of $R$ (respectively $f$) is $k$, then $F(R)$ (respectively $F(f)$) has $k$ free variables.

2. A formula $\delta$ of $\mathcal{L}_{\mathsf{T}_2}$, with one free variable, giving the domain of the interpretation.[3]

A well-known example of an interpretation is that of arithmetic into set theory, for instance of PA into ZFC — where the domain formula $\delta$ restricts the universe of sets to the finite ordinals. Then, for instance, the PA-constant $0$ can be translated as the empty set, the successor operation $S(x)$ as $x \cup \{x\}$, and so on.

A first minor adaptation that we make of Visser (1997)'s definition concerns variables. Visser extends $\mathcal{L}_{\mathsf{T}_2}$ with fresh variables in order to avoid variable clashes. Instead, we will make the simplified assumption that we can always map variables $x$ of $\mathcal{L}_{\mathsf{T}_1}$ to identically named variables in $\mathcal{L}_{\mathsf{T}_2}$ ($x \mapsto x$). In case this does cause variable clashes for a translated variable $x$, we will take $x$ to 'stand for' a suitable

[3]Note that we are here only allowing one-dimensional interpretations. This is not essential: surrogates can be made up of single elements or of tuples from the ontology of the interpreting theory. However, one-dimensionality provides for notational simplicity in further definitions and in the Appendix.

fresh variable of $\mathcal{L}_{\mathsf{T}_2}$. By keeping the variable names the same, we avoid focusing on purely formal aspects of interpretability, and keep its definition a bit more intuitive.[4]

Now $i$ gives our translation $(.)^i$ of $\mathcal{L}_{\mathsf{T}_1}$ into $\mathcal{L}_{\mathsf{T}_2}$ in the following way. First, $\bot$ is interpreted as itself. Further, the translation of a formula $\varphi$ will have the same free variables as $\varphi$ itself — while the translation of a term $t$ will have the free variables of $t$ plus one more fresh variable, standing for the *value* of $t$. However, as we send variables to themselves, term translations disappear when the term is a variable. In the definition, let $\overline{x_k}$ stand for a sequence of $k$ variable terms, and let $\overline{t_l}$ stand for a sequence of $l$ non-variable terms. (See Section 4.1 for the notation $\delta_{\overline{x}}$.)

- $R(\overline{t_l}, \overline{x_k})^i := \exists \overline{y_l}(\delta_{\overline{y_l}} \wedge F(R)(\overline{y_l}, \overline{x_k}) \wedge (t_1)^i(y_1) \wedge ... \wedge (t_l)^i(y_l))$
- $f(\overline{t_l}, \overline{x_k})^i := \exists \overline{y_l}(\delta_{\overline{y_l}} \wedge F(f)(\overline{y_l}, \overline{x_k}) \wedge (t_1)^i(y_1) \wedge ... \wedge (t_l)^i(y_l))$
- $(.)^i$ commutes with the propositional connectives
- $(\forall x \varphi)^i := \forall x(\delta(x) \to \varphi^i)$, $(\exists x \varphi)^i := \exists x(\delta(x) \wedge \varphi^i)$

The last item conveys that the quantifiers of translated formulas become relativized by the domain formula $\delta$. In the interpretation translation of predicates and function symbols (the first two items above), Visser introduces unrelativized existential quantifiers. As shown above, our second adaptation is that we relativize even these quantifiers. This will ensure that we can consistently restrict a proof to 'good' syntax later on.

Finally, we state the preservation of provability of an interpretation. $\mathsf{T}_2$ interprets $\mathsf{T}_1$ via $i$ if: for all theorems $\varphi$ of $\mathsf{T}_1$, $\vdash_{\mathsf{T}_2} \delta_\varphi \to \varphi^i$.

We also note that, given a model $\mathcal{M}$ of $\mathsf{T}_2$, an interpretation $i : \mathsf{T}_1 \to \mathsf{T}_2$ gives a way to induce a model of $\mathsf{T}_1$ on $\delta$ (Visser, 1997). Essentially, the objects of $\mathcal{M}$ are equated according to interpreted identity, after which the objects that satisfy $\delta$ are selected. Interpreted predicates and function symbols then work on these equivalence classes of objects (see for full details (Visser, 1997)). This can be seen as a formal way of making sense of surrogate content, analogous to how we can make sense of the ontology of a theory generally as a standard model.

### 4.3.2 Referring to extended content

This section will elaborate on how syntax restricted in a way inspired by interpretations can refer to surrogate and structural content. We discuss how interpretations can induce a natural notion of surrogate content of an ontology, and how they provide a way of preserving structural content.

[4] It will, however, have the consequence that we need to distinguish between variable and non-variable terms in the Appendix, Section 4.6, although this does not take too much effort.

**Referring to surrogate content**

The interpretation translation of $i : \mathsf{T}_1 \to \mathsf{T}_2$ gives a clear indication of how $\mathsf{T}_2$-syntax should refer to surrogate content. The domain formula $\delta$ of $i$ characterizes the collection of surrogate objects in $\mathsf{T}_2$. In particular, for any $\mathsf{T}_1$-formula $\varphi$ that indicates an object or operation in its ontology, $\varphi^i$ will indicate the surrogate version. And similarly to the transformation of a model of $\mathsf{T}_2$ into a model of $\mathsf{T}_1$ described above, we may view two objects in the ontology of $\mathsf{T}_2$ as exactly the same surrogate if they are equal under $F(=)$.

We suggest that the properties of the interpretation translation are suitable for inducing a notion of surrogate content as we informally mean it. An important feature of a translation is (a first-order variant of) *schematicity* as in (Incurvati and Nicolai, 2024), where "the translation of a complex formula [should be] a fixed schema of the translation of its parts". This is handled by the interpretation by its translation of a formula based on the individual translations of constants, function symbols and predicates occurring in it, and by its commutativity with propositional connectives. The translation is also injective, and the arity of function symbols and predicates is preserved, so that the translation in $\mathsf{T}_2$ makes the same distinctions between objects and operations as $\mathsf{T}_1$ does. This is confirmed by the fact that the interpretation translation allows for well-known definitions of surrogates in practice: for instance, both the Von Neumann ordinals and the Zermelo ordinals can easily be defined through $\delta$ (see for various other examples of interpretations (Visser, 1997)).

In addition to accepting the interpreted $\mathcal{L}_{\mathsf{T}_1}$-formulas, however, we are looking to characterize the part of the $\mathcal{L}_{\mathsf{T}_2}$-syntax that we can in practice restrict a $\mathsf{T}_2$-proof to, so that the restricted part of the proof only refers to surrogate content. This cannot simply be the set of interpreted $\mathcal{L}_{\mathsf{T}_1}$-formulas, as not all inference rules preserve 'being of interpreted form'. Instead, we propose to restrict a proof to certain 'good' $\mathcal{L}_{\mathsf{T}_2}$-formulas, of which the interpreted formulas will be a subset, and which comes down to an extension by instances of $\delta$ and by interpreted terms. Since instances of $\delta$ and interpreted terms only highlight specific elements within the surrogate ontology, this extension is harmless for referring to surrogate content. The definition of 'good formulas' is as follows.

**4.3.1 Definition (Good $\mathcal{L}_{\mathsf{T}_2}$-formulas).** Let $i : \mathsf{T}_1 \to \mathsf{T}_2$ be an interpretation with domain formula $\delta$. First, a $\mathcal{L}_{\mathsf{T}_2}$-term $t$ is *good* if it is either a variable, or a function symbol $f$ applied to terms $t_1, ..., t_n$ such that:

- Each $t_j$ is good ($1 \leq j \leq n$)
- $\vdash_{\mathsf{T}_2} \delta_{\bar{t}} \to \delta(f(\bar{t}))$

Now we define the set of $\mathcal{L}_{\mathsf{T}_2}$-formulas $\mathcal{S}$ by the following ingredients.

$$\mathcal{S} := \{\varphi^i \mid \varphi \in \mathcal{L}_{\mathsf{T}_1}\} \cup \{t^i \mid t \text{ a term of } \mathcal{L}_{\mathsf{T}_1}\} \cup \{\delta(t) \mid t \text{ a good term of } \mathcal{L}_{\mathsf{T}_2}\}$$

We then define the *good* $\mathcal{L}_{\mathsf{T}_2}$-formulas as follows:

- Each $\varphi \in \mathcal{S}$ is good
- If $\varphi$ and $\psi$ are good, then $\varphi \circ \psi$ is good ($\circ \in \{\wedge, \vee, \rightarrow\}$)
- If $\varphi$ is good, $\exists x(\delta(x) \wedge \varphi)$ is good
- If $\varphi$ is good, $\forall x(\delta(x) \rightarrow \varphi)$ is good

Thus, the set of good $\mathcal{L}_{\mathsf{T}_2}$-formulas will be exactly what we will restrict a $\mathsf{T}_2$-proof to for secondary ontological purity. Finally, we note here that Arana (2017) has argued that interpretations are not sufficient to preserve topical purity, as translations do not preserve understanding. Specifically, Arana rejects the idea that "if two theories $\mathsf{T}_1$ and $\mathsf{T}_2$ are mutually interpretable, then their semantic parts (terms, statements) have identical meanings", because then topical purity does not capture mathematical practice anymore. We agree with both points, and emphasize that we only claim a correspondence between $\mathcal{L}_{\mathsf{T}_1}$-syntax and the set of good $\mathcal{L}_{\mathsf{T}_2}$-formulas. Nor do we claim that these two sets have a fully identical ontology, but we will argue in the next section that they have an ante rem structure in common, which suffices for our sense of secondary ontological purity.

**Referring to structural content**

We see a theory as referring to an ante rem structure through the reference to its ontology. Thus, each formula $\varphi$ describing an object in the ontology $\mathsf{T}_1$ can be seen to additionally refer to the structural place underlying this object in the ante rem structure. It may differ per ontology what relations the structure preserves (in 'placeholder' form). Given the relations of a structure, however, their ontological versions will certainly be captured by the formulas of $\mathsf{T}_1$. If anything, the structure will have less (detailed) properties than a full ontology, so that $\mathsf{T}_1$ should always be able to refer (by an ontological instance) to what the structure is made up of.[5]

The preservation of provability of interpretations now ensures that reference to this structure is preserved by the interpreted formulas in $\mathsf{T}_2$. That is, for each $\mathcal{L}_{\mathsf{T}_1}$-formula $\varphi$ that refers to an element of its ante-rem structure through its ontology, $\varphi^i$ in $\mathcal{L}_{\mathsf{T}_2}$ can be seen to refer to the same element of the structure through its surrogate ontology. If we take $\mathsf{T}_1$ to refer to some ontology, and $\mathsf{T}_2$ to be able to refer to surrogates of this ontology, then we should also accept that both of them can refer to what these ontologies have in common. Preservation of provability makes exactly the right connection between an ontology and a surrogate ontology,

[5] Shapiro (1997) notes that the natural numbers with just a successor operation, or the natural numbers with, e.g., additionally an order relation, ideally describe the same structure. Here, we only claim a structure is preserved when all provable properties of a theory are preserved; and so we distinguish more structures than Shapiro ultimately intends. Secondary purity could thus be extended even further — but for now, we have at least ensured structure preservation.

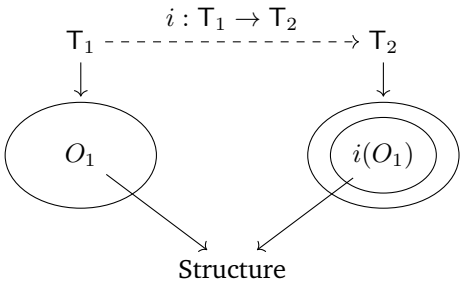



Figure 4.1: A visual representation of the reference of theories $\mathsf{T}_1$ and $\mathsf{T}_2$ to the ontologies $O_1$ and $O_2$, and of the structure underlying $O_1$ and the surrogate ontology $i(O_1)$.

and the ante rem structure that both of them instantiate. The general situation is illustrated in Figure 4.1.

### 4.3.3 A restriction on proof rules

We are now set to define the two ingredients that will culminate in the derivation criterion in the next section, restricting formal proofs to 'good' formulas. The first ingredient selects particular instances of 'good' formulas (the interpreted $\mathsf{T}_1$-axioms, $\delta$-instances, and instances of interpreted functionality and identity axioms). Considering each natural deduction proof as a tree $(V, E)$ as in Section 4.1, part of the derivation criterion will be to require that one of these formulas occurs in each branch. Here, we will write $=^i$ for interpreted equality, and use $\overline{x} =^i \overline{y}$ for the conjunction $x_1 =^i y_1 \wedge ... \wedge x_n =^i y_n$.

**4.3.2 Definition (Pure $\mathcal{L}_{\mathsf{T}_2}$-formulas).** Let $i : \mathsf{T}_1 \to \mathsf{T}_2$ be an interpretation with domain formula $\delta$. Then a $\mathcal{L}_{\mathsf{T}_2}$-formula is *pure* if either:

1. $\varphi$ is an interpreted (non-)logical axiom of $\mathsf{T}_1$.
2. $\varphi = \delta(t)$, for $t$ a good term of $\mathcal{L}_{\mathsf{T}_2}$.
3. $\varphi$ is a relativized uniqueness or totality axiom for interpreted function symbols. This means that $\varphi$ can be:
   (a) *(Uniqueness)* For any $\mathcal{L}_{\mathsf{T}_1}$-term $t$,
   $$\forall x \forall y(\delta(x) \wedge \delta(y) \to (t^i(x) \wedge t^i(y) \to x =^i y)))^6$$
   (b) *(Totality)* For any $\mathcal{L}_{\mathsf{T}_1}$-term $t$,

[6]The usual definition of Uniqueness uses an implication between $\delta(x)$ and $\delta(y)$. Here we use the equivalent definition by conjunction, so that we can use the abbreviation $\delta_{\overline{x}}$ in the Appendix in long natural deduction proofs. The same holds for the substitution axioms in 4.

$$\exists y(\delta(y) \wedge t^i(y))$$

4. $\varphi$ is a relativized first-order equality axiom for interpreted equality.[7] This means that $\varphi$ can be:

   (a) *(Reflexivity)*

   $$\forall x(\delta(x) \to x =^i x)$$

   (b) *(Function substitution)* For any $\mathcal{L}_{\mathsf{T}_1}$-function symbol $f(\overline{x})$,

   $$\forall \overline{x}\forall \overline{y}(\delta_{\overline{x}} \wedge \delta_{\overline{y}} \to (\overline{x} =^i \overline{y} \to (F(f)(\overline{x}, z) \to F(f)(\overline{y}, z))))$$

   (c) *(Formula substitution)* For any $\mathcal{L}_{\mathsf{T}_1}$-formula $\varphi$,

   $$\forall \overline{x}\forall \overline{y}(\delta_{\overline{x}} \wedge (\delta_{\overline{y}} \to (\overline{x} =^i \overline{y} \to (\varphi^i(\overline{x}) \to \varphi^i(\overline{y})))))$$

The second ingredient specifies which applications of the inference rules of the natural deduction proof system preserve 'goodness' of formulas. Thus, we are not defining a new proof system, but we emphasize for a rule $R$ to which instances (denoted by $R^i$) it should be restricted in proofs that are to refer to surrogate (and structural) content.

**4.3.3 Definition** (**Pure rule applications**)**.** We provide four restricted inference rules, so that if the premises of the rule are good, then the conclusion of the rule is good as well. It is easily verifiable that the other remaining rules already preserve goodness — we will call their instances together with those of the restricted rules *pure rule applications*.

- *Disjunction introduction*. In order to ensure that the conclusion of this rule is good, we require that the introduced disjunct is also good (denoted by $\psi_g$).

$$\frac{\varphi}{\varphi \vee \psi} \vee I \qquad \frac{\varphi}{\varphi \vee \psi_g} \vee I^i$$

- *Universal introduction*. In order to ensure that the conclusion of this rule is good, we require that the form of the premise must be restricted to implications with antecedent $\delta$.

$$\frac{\varphi(y)}{\forall x\varphi(x)} \forall I \qquad \frac{\delta(y) \to \varphi(y)}{\forall x(\delta(x) \to \varphi(x))} \forall I^i$$

- *Existential introduction*. In order to ensure that the conclusion of this rule is good, we require that the form of the premise is restricted to conjunctions, with $\delta$ as one of the conjuncts.

[7] Note that the equality axioms technically fall under 1 as the interpretation of a logical $\mathsf{T}_1$-axiom. We add them for clarity, as they will be used regularly in the Appendix.

$$\frac{\varphi[x/t]}{\exists x\varphi(x)}\,\exists I \qquad \frac{(\delta \wedge \varphi)[x/t]}{\exists x(\delta(x) \wedge \varphi(x))}\,\exists I^i$$

- *Universal elimination.* In order to ensure that the conclusion of this rule is good, we require that the instantiation with term $t$ in the conclusion is such that $t$ is good (denoted by $t_g$).

$$\frac{\forall x\varphi(x)}{\varphi[x/t]}\,\forall E \qquad \frac{\forall x\varphi(x)}{\varphi[x/t_g]}\,\forall E^i$$

## 4.4 Criteria for secondary ontological (im)purity

We combine the previous definitions and state various criteria for secondarily ontological (im)purity.

### 4.4.1 Secondary ontological purity of formal proofs

First, consider a criterion for when a formal proof is secondarily ontologically pure.

**4.4.1 Definition (Criterion for secondary ontological purity of formal proofs).** Suppose we are given an informal theorem corresponding to the ontological context $(O, \varphi_{\mathsf{T}_1}, R)$. Suppose we are given a natural deduction proof in $\mathsf{T}_2$ of $\varphi$, where $\varphi$ is a formalization of the theorem into $\mathcal{L}_{\mathsf{T}_2}$. Let $\varphi$ instantiate the node $v$ in the corresponding tree $(V, E)$. Then the proof is *secondarily ontologically pure* (denoted $\Gamma \vdash^{P}_{\mathsf{T}_2} \varphi$) if there exists an interpretation $i : \mathsf{T}_1 \to \mathsf{T}_2$ such that:

- Any assumption instantiating a node $a$ in the proof is *good*, and the open branch $aE...Ev$ contains only *pure rule applications*.
- Any closed branch contains a node $p$ that is instantiated by a *pure formula*, such that the final branch part $pE...Ev$ contains only *pure rule applications*.

This criterion ensures that, starting from the leaves, at some point in the proof each branch is restricted to good formulas only. We will sometimes refer to this criterion as 'the derivation criterion' for conciseness.

#### Robustness of the criterion

In order to show that this criterion is robust, we show that the criterion has instances. Namely, we show that $\mathsf{T}_2$ has a secondarily ontologically pure formal proof of every interpreted theorem of $\mathsf{T}_1$. For this, we use what we call 'simulation'. Recall that $\delta_{\lambda_{\mathcal{D}}}$ is the conjunction of $\delta$'s applied to all the free variables of formulas occurring in the proof $\mathcal{D}$. The instances of $\delta$ are needed to guarantee provability of interpreted formulas (see Section 4.3.1). Furthermore, let $\Gamma^i$ be the set $\{\gamma^i \mid \gamma \in \Gamma\}$.

**4.4.2 Definition** (**$\mathsf{T}_1$-simulation**). Let $i : \mathsf{T}_1 \to \mathsf{T}_2$ be an interpretation. Let $\mathcal{D}$ refer to $\Gamma \vdash_{\mathsf{T}_1} \varphi$, and let it correspond to a tree $(V, E)$. Then a *simulation* of $\mathcal{D}$ in $\mathsf{T}_2$ is a proof $\Gamma^i, \delta_{\lambda_\mathcal{D}} \vdash_{\mathsf{T}_2} \varphi^i$, corresponding to $(V', E')$, with the following requirements.

- For every branch step $v_{n+1}Ev_n$ in $(V, E)$ such that $\chi$ instantiates $v_{n+1}$ and $\psi$ instantiates $v_n$, there exists a sequence $v_mE...Ev_k$ $(m > k)$ in $(V', E')$ such that $\chi^i$ instantiates $v_m$ and $\psi^i$ instantiates $v_k$.
- Take any sequence of nodes $v_{n+2}Ev_{n+1}Ev_n$ in $(V, E)$. Suppose $v_{n+2}Ev_{n+1}$ corresponds to the sequence $v_mE...Ev_k$ in $(V', E')$ $(m > k)$, and $v_{n+1}Ev_n$ corresponds to the sequence $v_lE...Ev_q$ in $(V', E')$ $(l > q)$. Then $v_k = v_l$.

Intuitively, this is a simulation, as each inference step is replaced by a proof from its interpreted premises to the interpreted conclusion. More formulas can be added in between the interpreted formulas, however, to secure provability. Thus, there are various ways of simulating a proof. We now consider the theorem, which shows there are *pure* simulations.

**4.4.3 Theorem** (**Existence of pure simulations**). *Let $i : \mathsf{T}_1 \to \mathsf{T}_2$ be an interpretation. Let $\mathcal{D}$ be the proof $\Gamma \vdash_{\mathsf{T}_1} \varphi$ in the classical first-order natural deduction calculus. Then $\Gamma^i, \delta_{\lambda_\mathcal{D}} \vdash^P_{\mathsf{T}_2} \varphi^i$ by simulation.*[8]

*Proof.* The proof idea is as follows: we provide the pure $\mathsf{T}_2$-simulations of each inference rule in $\mathsf{T}_1$. Then a complete $\mathsf{T}_1$-proof is simulated in $\mathsf{T}_2$ by pasting together the simulations of the individual rule applications. For propositional rules, and the rules $\forall$I and $\exists$E, it is easy to see there are pure simulations of any use of the rule. For $\forall$E and $\exists$I, we need a lemma to deal with interpreted formulas of the form $(\varphi(t))^i$, which cannot easily be rewritten, as the term translation $t^i$ appears only at the atomic level of translations of predicates and function symbols inside $\varphi$. To provide a simulation of $\forall$E and $\exists$I, then, we take a detour through the formula $\exists x(\delta(x) \wedge t^i(x) \wedge \varphi^i(x))$. The full proof can be found in the Appendix (Section 4.6). □

**A (very) small working example**

Take the interpretation $i : \mathsf{Q} \to \mathsf{C}^2_{\mathsf{FO}}$ of Robinson's Q into a theory of concatenation that has signature $(*, a, b)$. The interpretation is defined in (Ganea, 2009). We provide the domain formula and the interpretation of the constant zero (which suffices for the example). The interpretation is identity-preserving.

[8] Although it is very likely that the other direction of this theorem also holds (if $\varphi^i$ can be proven in $T_2$ in a pure way, then $T_1$ proves $\varphi$), we do not show it here. Note that, although this is a desirable property, a counterexample would merely show that a notion can be made sense of in terms of surrogate ontology, in a more complex way than $T_1$ finds acceptable.

- $\delta(x) := T(a,x) \vee x = b$. Here, $T(a,x)$ is an abbreviation for saying $x$ is a string made up entirely of $a$'s. Thus, a natural number $n > 0$ is represented in $\mathsf{C}^2_{\mathrm{FO}}$ as a string of $a$'s of length $n$.
- $0^i := x = b$.

Now take the very simple Q-proof:

$$\dfrac{\forall x(x = x)}{0 = 0}\,\forall\mathrm{E}$$

And consider a similar proof in $\mathsf{C}^2_{\mathrm{FO}}$.

$$\dfrac{\dfrac{\begin{array}{c}\mathsf{C}^2_{\mathrm{FO}}\\ \vdots\\ \forall\mathbf{x}((\mathbf{T}(\mathbf{a},\mathbf{x}) \vee \mathbf{x} = \mathbf{b}) \to \mathbf{x} = \mathbf{x})\end{array}}{(T(a,b) \vee b = b) \to b = b}\,\forall\mathrm{E} \qquad \dfrac{\dfrac{\forall x(x = x)}{b = b}\,\forall\mathrm{E}}{\mathbf{T}(\mathbf{a},\mathbf{b}) \vee \mathbf{b} = \mathbf{b}}\,\vee\mathrm{I}}{b = b}\,\to\mathrm{E}$$

Marked bold are the pure formulas occurring in each branch: from that moment on, the proof is secondarily ontologically pure. Note that this proof satisfies the derivation criterion, but is *not* a simulation, as we reach $b = b$ instead of the literal translation $(0 = 0)^i$. The Appendix (Section 4.6) provides a way to properly simulate $\forall$E that does end at $(0 = 0)^i$, but it is one of the tedious cases of simulation, and not helpful for intuitions in an example. Still, $b = b$ is a natural translation of $0 = 0$, and its proof as shown here is also an intuitive imitation of $\forall$E. This suggests that the notion of interpretation, and so that of simulation, could be extended to include different translations, in case the languages $\mathcal{L}_{\mathsf{T}_1}$ and $\mathcal{L}_{\mathsf{T}_2}$ are quite alike (e.g., where constants can be interpreted directly as constants).

**Remarks on the criterion**

We here discuss some aspects of the derivation criterion worth mentioning. First, it matches the notion of equivalence for context theories well. Suppose $\mathsf{T}'$ definitionally extends $\mathsf{T}$ by a symbol $S$ defined by $\varphi$. Suppose also that we have an interpretation $i_1 : \mathsf{T} \to \mathsf{V}$. Then we can naturally define $i_2 : \mathsf{T}' \to \mathsf{V}$ by setting $S^{i_2} = \varphi^{i_2} = \varphi^{i_1}$. Note that the definitional axiom of $\mathsf{T}'$ then becomes an interpreted tautology in $\mathsf{V}$ of the form $\forall\overline{x}(\delta_{\overline{x}} \to ((\varphi(\overline{x}))^{i_2} \leftrightarrow (\varphi(\overline{x}))^{i_2}))$. And, as $\varphi^{i_1} = \varphi^{i_2}$, this is also an interpreted tautology with respect to $i_1$. This means that the derivation criterion resulting from $i_2$ is indistinguishable from the one resulting from $i_1$, and shows some consistency between full ontological purity and secondary ontological purity: if we consider two theories to be equivalent as context theories for a certain theorem, then we see that they can naturally induce the same derivation criterion for purity of formal proofs in a third theory.[9]

[9]Note, however, that there may be choices of $S^i$ that do not exactly coincide with $\varphi^i$. For example, we could add to PA a symbol $E(x)$ that says that $x$ is even. In PA, the defining formula could be

Furthermore, we remark that any formal proof that satisfies the derivation criterion for an interpretation $i : \mathsf{T} \to \mathsf{V}$ needs to prove the pure formulas from $\mathsf{V}$ itself. Additionally, strictly speaking, $\mathcal{L}_{\mathsf{V}}$-quantifiers (even when relativized by $\delta$) should be taken as ranging over the entire $\mathsf{V}$-domain. Thus, the pure formulas of $\mathcal{L}_{\mathsf{T}}$ cannot refer exclusively to the surrogate content — rather, they illuminate or highlight the surrogate content within the content of $\mathsf{V}$. Ideally, perhaps, we would like to have formal proofs of $\mathsf{V}$ refer to nothing but the surrogate content, to fully eliminate any extraneousness. However, the embedding in the content of $\mathsf{V}$ may well be the only way in which surrogate content is given meaning. Namely, the 'pure formulas' are built from primitives whose ontological meaning is properly given by the axioms of $\mathsf{V}$; and it is unclear whether they can be thought of as independently corresponding to the surrogate ontology (without the connection to $\mathsf{V}$). On a slightly different note, because formal proofs satisfying the derivation criterion start with the fully powered $\mathsf{V}$-axioms, they are allowed to refer to extraneous content before the derivation of the pure formulas. When calling these proofs pure, then, we restrict that statement to the part of such proofs that comes after the pure formulas. Still, perhaps we may also view the part before the pure formulas as clarifying the way in which the interpreting theory approaches the subject matter of the context theory, and showing how this material fits into the interpreting theory.

### 4.4.2 Secondary ontological purity for informal proofs

We presented an elaborate criterion that characterizes which formal proofs are secondarily ontologically pure. It tells us that any interpretation induces a sense of surrogate content and secondarily ontologically pure formal proofs. Like formal proofs, an informal proof should be secondarily ontologically pure if it only draws on the surrogate ontology of the theorem. As before, this will require a connection between the notions in an informal proof, and their interpretation in a surrogate ontology. Given an ontological context $(O, \varphi_{\mathsf{T}_1}, R)$ for a theorem, we may judge this by checking *whether* the notions in a proof are formalizable in $\mathsf{T}_1$ or equivalently, given an interpretation $i : \mathsf{T}_1 \to \mathsf{T}_2$, *whether* they are formalizable in terms of the good formulas of $\mathcal{L}_{\mathsf{T}_2}$. Since the notions in the informal proof can intuitively concern a very different ontology than $O$, the formalization in $\mathsf{T}_1$ does not anymore have to be 'natural'. We do require that there exists a formalization in $\mathsf{T}_2$ that is natural, so that $\mathsf{T}_2$ refers to the right ontology associated to the notions in the proof. However, given an interpretation, the formalization in terms of the good formulas in $\mathsf{T}_2$ does *not* need to be natural (although of course it can be). This is because the good formulas will code the notions of an informal proof in a way that $\mathsf{T}_2$ may not itself.

$\varphi_1(x) = \exists y(x = y + y)$, or $\varphi_2(x) = \exists y(x = SS0 \cdot y)$. When a theory interprets PA by $i$, it is clear that the translations $\varphi_1^i$ and $\varphi_2^i$ will not be the same. Hence, one could theoretically pick $\varphi_1$ to define the definitional extension of PA, but pick $\varphi_2^i$ as the interpreting translation of $E(x)$.

Still, we will know that the informal proof is concerned with the ontology of $\mathsf{T}_2$, and that it can be made sense of in terms of $\mathsf{T}_1$-surrogates; a version of the pure content. That is, we just need to know, given an informal proof, that it is possible to make sense of its notions in terms of the restricted surrogate ontology. If this is the case, secondary ontological purity holds. Thus, we will maintain the following definition.

**4.4.4 Definition (Criterion for secondary ontological purity of informal proofs).** Given an informal mathematical theorem corresponding to the ontological context $(O, \varphi_{\mathsf{T}_1}, R)$, an informal proof of the theorem is *secondarily ontologically pure* if there exists a formalization of any notion in the proof in $\mathsf{T}_1$, if there exists a *natural* formalization of any notion in the proof in some theory $\mathsf{T}_2$, and if there exists an interpretation $i : \mathsf{T}_1 \to \mathsf{T}_2$.

**4.4.5 Example (Infinitude of Primes).** Consider again the arithmetical proof of IP, which we have established is fully ontologically pure. Trivially, this will be secondarily ontologically pure with respect to PA, for the identity interpretation $i : \mathsf{PA} \to \mathsf{PA}$. Thus, full ontological purity implies secondary ontological purity.

Now consider Furstenberg's topological proof of IP, which we established before as not fully ontologically pure. The notions used in the informal proof arguably correspond to an ontology of sets, which are used to construct a topology, arithmetic sequence, and even the integers and natural numbers. Thus, a suitable set theory could provide a natural formalization of these notions, say ZFC. The context theory for IP remains PA as before, and as argued in Section 3.4.2, Furstenberg's proof has a formalization in PA (although not a natural one). Now it is easy to come up with an interpretation $i : \mathsf{PA} \to \mathsf{ZFC}$, where for instance $\delta(x) := x \in \omega$. As the topology used in IP is reducible to an ontology of natural numbers (recall the considerations of Example 3.4.4), it is also reducible to the surrogate ontology of set-theoretic numbers — and since a set-theoretic ontology is associated to the notions in the proof, this establishes secondary ontological purity.

By using interpretability results, secondary ontological purity results broaden our understanding of how disciplines of mathematics represent each other. They encourage us to see connections between formalizations of the same informal proof in different theories, and even within a theory. And, they reassure us that the notions we use in a proof have an ontologically (though not necessarily epistemically) harmless formalization. Instead of separating, for instance, the arithmetical and the topological informal proofs of IP, a secondary ontological purity result encourages us to delve more into their connections. Aside from ontological similarities, we may continue the comparison by looking at their proof strategies, such as in (Carlson, 2014), and possibly connect these findings again to epistemic properties of the proofs.

### 4.4.3 Interaction between full and secondary ontological purity

We shortly discuss the interaction between the two types of purity. For formal proofs, it should be noted that fully ontologically pure and secondarily ontologically pure proofs of the same theorem intersect if a proof satisfies the derivation criterion within the context theory or a definitional extension. However, there are also fully ontologically pure proofs that do not satisfy the derivation criterion, as well as secondarily ontologically pure proofs that are not proven in the context theory or a definitional extension. Thus, the criteria for full and secondary ontological purity are independent, but may happen to overlap.

For informal proofs, we saw that full ontological purity always implies secondary ontological purity, as guaranteed by the identity interpretation from and to the context theory. The informal side just asks whether the theory under consideration naturally formalizes the proof, and subsequently asks for the existence of a formalization of the proof into the context theory and an interpretation of the context theory. The latter is just a box to check, after which we know the proof is formalizable in terms of a sense of surrogate content. Then an informal proof is secondarily ontologically pure. We see that this differs on the formal side, because we can there more easily distinguish between different types of surrogate content, and emphasize that it is the specific formalization in terms of the surrogate content that is secondarily ontologically pure. However, just like on the formal side, there also are secondarily ontologically pure proofs that are not fully ontologically pure, if they are not naturally formalizable into the context theory.

### 4.4.4 Impurity of proof

Lastly, based on the criteria for purity that we have formulated so far, this section mentions some criteria for impurity of proof. Secondary impurity results show that there are ingredients of a proof that cannot be, in an acceptable way, represented by an ontology. Full impurity results show that a proof really properly leaves a (surrogate) ontology, and must go beyond representing the context theory.

Full ontological purity is based on a context theory (or multiple), definitional extensions and their ontology. Its negation gives rise to a secondary sense of impurity.

**4.4.6 Definition** (**Criteria for secondary impurity**)**.** Suppose we are given an ontological context $(O, \varphi_{\mathsf{T}_1}, R)$ for an informal theorem. An informal proof is *secondarily impure* if some notion $A$ in the proof is not naturally formalizable in $\mathsf{T}_1$. A formal proof $\Gamma \vdash_{\mathsf{T}_2} \varphi_{\mathsf{T}_2}$ is *secondarily impure* if $\mathsf{T}_2 \notin [\mathsf{T}_1]$.

Thus, secondary impurity tells us that the proof does not concern itself with the true ontological content of a theorem, but that it still might be representable by a surrogate version of this content. Note that the condition for formal proofs relies on the idea that $[\mathsf{T}_1]$ contains *all* theories that capture the right ontology.

This will not always be the case. Thus, secondary ontological impurity for formal proofs simply says we do not have the purity guarantee.

For full impurity, suppose again we have an ontological context $(O, \varphi_{\mathsf{T}_1}, R)$ for an informal theorem. Full impurity of informal proofs then negates the main aspect of secondary ontological purity.

**4.4.7 Definition** (**Criterion for full impurity of informal proofs**)**.** An informal proof is *fully impure* if some notion in the proof cannot be formalized in $\mathsf{T}_1$.

Thus, full impurity tells us the ontology of the theorem is simply insufficient for capturing a proof, and subsequently any surrogate ontology is, too. Now for a formal proof, we could say that it is fully impure if it does not satisfy the derivation criterion for any interpretation. But this does not fully capture it: while the derivation criterion guarantees purity, it does not exclude other pure proofs from existing. Thus, an all-encompassing guarantee for full impurity is hard to specify; it requires us to figure out when an interpreting theory uses its full strength 'only to reach the interpreted theory', but not to do anything meaningful in the proof itself. Instead, we give one case that we can make clear.

**4.4.8 Definition** (**Criterion for full impurity of formal proofs**)**.** Suppose we are given an informal theorem corresponding to the ontological context $(O, \varphi_{\mathsf{T}_1}, R)$. Let $i : \mathsf{T}_1 \to \mathsf{T}_2$ be an interpretation. Then a natural deduction proof in $\mathsf{T}_2$ of $\varphi^i_{\mathsf{T}_1}$, where $\varphi^i_{\mathsf{T}_1}$ instantiates the node $v$ in the corresponding tree $(V, E)$, is *fully impure* if:

- Any assumption instantiating a node $a$ in an open branch is *good*.
- Any closed branch in $(V, E)$ contains a node $p$ instantiated by a *pure formula*.
- There is a final branch part $aE...Ev$ or $pE...Ev$ that uses an inference rule which violates the restriction of Definition 4.3.3.

This says that at some point in the proof, we do reach the point where we restrict to the interpreted theory. However, by violating one of the restrictions on proof rules, we leave this interpretation again during the proof, and thus leave a description of surrogate content by our notion of good formulas. We finish this section with two examples of informal proofs.

**4.4.9 Example** (**Planar Desargues's Theorem**)**.** Consider Planar Desargues's Theorem, with an ontology of planar geometrical notions (such as points and lines), and let the spatial axioms of Hilbert's incidence and order axioms (Group I and II, see (Hallett, 2007)) be the context theory (of course, other choices can be made). The argument here is simple, although it should be noted that it uses the unproven but highly likely other direction of Theorem 4.4.3, as mentioned in footnote 17: as the theorem is unprovable in the context theory, we expect it is unprovable in

any other theory that interprets the context theory, when this theory is restricted to just the surrogate content. This would mean that any informal proof of the theorem that we do have, is not formalizable in the context theory, and is necessarily fully impure.

**4.4.10 Example** (**Infinitude of Primes**)**.** Consider the topological proof of IP, which we saw before has secondary ontological purity. However, we see here that it also has secondary impurity, as a consequence of the result in Section 3.4.2 that it is not fully ontologically pure: we explained in Example 3.4.4 why we do not think that there is a natural formalization of all the notions occurring in the proof into PA. This emphasizes that 'secondary ontological purity' really lowers the purity level, as it can co-occur with an unnatural ontology.

Thus, this nuances the 'pure/impure' distinction: we suggest there is full ontological purity, full impurity, and a level in between that exhibits properties of both purity and impurity. The incorporation of other notions of purity should be able to induce even more levels of purity on top of these.

## 4.5 Conclusion

In this chapter, we have supplied formal and informal proofs with a notion of secondary ontological purity, based on surrogate ontological content and structural content. For formal proofs, we suggest that the satisfaction of a derivation criterion motivated by an interpretation of the context theory into another theory guarantees secondary ontological purity. Namely, the interpretation criterion guarantees preservation of the structural content underlying surrogate content. For informal proofs, formalizability in the context theory together with natural formalizability into an interpreting theory suffices: i.e., even though the proof has a different natural ontology than the context theory, it is still formalizable into the context theory, allowing exactly surrogate definitions. Secondary ontological purity induces a more general notion of purity than traditional conceptions, which matches more a structuralist way of thinking, and encourages a more nuanced distinction between levels of purity.

Future research can take several directions. For one, the formal guarantee for secondary ontological purity relies on the idea that the full power of a theory is only used to 'get to' the interpreted theory, and not for any notion essential to the proof. For this, our approach relies on the order of inference steps in natural deduction proofs. However, there may be other ways of determining whether 'extraneousness' of a theory is relevant in a meaningful way. This would lead to more encompassing guarantees for formal (im)purity.

Furthermore, there are several possibilities for extending the approach of this paper. For instance, instead of restricting ourselves to first-order mathematical

theories, one might instead want to take into account second- or higher-order theories, which may also provide a natural counterpart for ontological content. Additionally, extensions of the derivation criterion may be found. An interpretation consists of a domain relativization and a translation function — but for theories that already have a very similar domain to the context theory, it would be more elegant to leave out a (trivial) relativization (e.g., when $\delta(x) := x = x$). Alternatively, when a theory differs only from the context theory in that it captures more objects (of a similar ontology), a domain relativization might be all that we need. This might also give rise to variants of the derivation criterion that induce 'domain purity' and 'operational purity', instead of purity in both aspects. A related point is that we might consider how other translations, aside from the interpretation translation, lead to proof simulations. In particular, we might look for translations that induce syntax restrictions that preserve other valuable parts of ontological content, generating new variants of ontological purity.

Finally, in the spirit of Chapter 2, we may further investigate the relation between (full and secondary) ontological purity, and various ideals of proof. Besides explanation, it has for instance been suggested previously that impurity relates to simplicity of proof (Arana, 2017; Iemhoff, 2017).

The case study of purity of Chapter 3 and 4 thus presents a method for analyzing an ideal of proof in the setting of formal proofs. We may think of formal proofs satisfying a criterion for full or secondary ontological purity, as potential formal counterparts of informal proofs that possess ontological purity — preserving one particular aspect of both types of proofs. However, at least in our approach, subjective interpretation will always play a role in establishing the link between informal and formal proofs.

## 4.6 Appendix: a proof translation

We here show rigorously by induction on the length of a derivation that, given an interpretation $i : \mathsf{T}_1 \to \mathsf{T}_2$, any proof in $\mathsf{T}_1$ can be simulated in a pure way in $\mathsf{T}_2$:

**4.6.1 Theorem.** *Let $i : T_1 \to T_2$ be an interpretation. Let $\mathcal{D}$ be the proof $\Gamma \vdash_{T_1} \varphi$ in the classical first-order natural deduction calculus. Then $\Gamma^i, \delta_{\lambda_\mathcal{D}} \vdash^P_{T_2} \varphi^i$ by simulation.*

We will use the notation $(\varphi(t))^i$ when $t$ is a non-variable term, and the notation $\varphi^i(t)$ when $t$ is a variable. This emphasizes that, in the first case, $t$ itself needs to be translated into a separate formula, while in the second case, we will map $t$ to $t$ itself. Furthermore, in order to make big natural deduction proofs readable we sometimes split them up: asterisks $*_j$ will indicate that this spot needs to be filled by the formula labeled by $(*_j)$, attaching two proofs together. Symbols $\approx_j$ will indicate that this spot needs the formula labeled by $(\approx_j)$ (and its proof) to be

repeated there.

Before embarking on the proof of the full theorem, however, we need two lemmas for the case where $\chi$ was obtained by an application of $\forall$E, and where it was obtained by $\exists$I. These cases involve formulas $(\varphi(t))^i$, which cannot easily be manipulated since $t^i$ occurs only at the depth of atomic formulas. We distinguish between two cases: for non-variable terms, we take a detour here for their simulation through $\exists x(\delta(x) \wedge \varphi^i(x) \wedge t^i(x))$. The case of variable terms is included at the end of the full proof, which is easier as it has no term translations.

**4.6.2 Lemma.** *Given an interpretation $i : T_1 \to T_2$, we have the following proofs for any $T_1$-formula $\varphi$ and non-variable term $t$:*

$$(\forall x \varphi(x))^i \vdash^P_{T_2} \exists x(\delta(x) \wedge \varphi^i(x) \wedge t^i(x))$$
$$\exists x(\delta(x) \wedge \varphi^i(x) \wedge t^i(x)) \vdash^P_{T_2} (\exists x \varphi(x))^i$$

*Proof.* The first is given by the following derivation, which uses totality. Let assumption [1] be $[\delta(y) \wedge t^i(y)]$.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[\forall x(\delta(x) \to \varphi^i(x))]}{\delta(y) \to \varphi^i(y)}\forall\mathrm{E} \quad \dfrac{[1]}{\delta(y)}\wedge\mathrm{E}}{\varphi^i(y)}\to\mathrm{E} \quad \dfrac{[1]}{t^i(y)}\wedge\mathrm{E}}{\varphi^i(y) \wedge t^i(y)}\wedge\mathrm{I} \quad \dfrac{[1]}{\delta(y)}\wedge\mathrm{E}}{\delta(y) \wedge \varphi^i(y) \wedge t^i(y)}\wedge\mathrm{I}}{\exists x(\delta(x) \wedge \varphi^i(x) \wedge t^i(x))}\exists\mathrm{I} \qquad \exists x(\delta(x) \wedge t^i(x))}{\exists x(\delta(x) \wedge \varphi^i(x) \wedge t^i(x))}\exists\mathrm{E}^1$$

The second proof is given by the derivation below. We use totality, uniqueness, and an equality axiom. Let [1] be $[\delta(w) \wedge t^i(w)]$, and let [2] be $[\delta(z) \wedge \varphi^i(z) \wedge t^i(z)]$.

$$\dfrac{\dfrac{\dfrac{\dfrac{[2]}{t^i(z)}\wedge\mathrm{E} \quad \dfrac{[1]}{t^i(w)}\wedge\mathrm{E}}{t^i_{z,w}}\wedge\mathrm{I} \qquad \dfrac{\dfrac{[2]}{\delta(z)} \quad \dfrac{[1]}{\delta(w)}}{\delta_{z,w}\ (\approx)} \qquad \dfrac{\forall xy(\delta_{x,y} \to (t^i_{x,y} \to x =^i y))}{\delta_{z,w} \to (t^i_{z,w} \to z =^i w)}\forall\mathrm{E}}{t^i_{z,w} \to z =^i w}\to\mathrm{E}}{z =^i w\ (*_1)}$$

$$\dfrac{\dfrac{\dfrac{\forall xy(\delta_{x,y} \to (x =^i y \to (\varphi^i(x) \to \varphi^i(y))))}{\delta_{z,w} \to (z =^i w \to (\varphi^i(z) \to \varphi^i(w)))}\forall\mathrm{E} \qquad \approx}{z =^i w \to (\varphi^i(z) \to \varphi^i(w))} \qquad *_1}{\varphi^i(z) \to \varphi^i(w)\ (*_2)}\to\mathrm{E}^*$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[2]}{\varphi^i(z)}\,\wedge\text{E}}{\varphi^i(w)}\,{}_{*_2}\!\!\to\text{E} \quad \dfrac{[1]}{\delta(w)}\,\wedge\text{E}}{\delta(w)\wedge\varphi^i(w)}\,\wedge\text{I}}{\exists x(\delta(x)\wedge\varphi^i(x))}\,\exists\text{I} \qquad \exists x(\delta(x)\wedge t^i(x))}{\exists x(\delta(x)\wedge\varphi^i(x))}\,\exists\text{E}^1 \qquad [\exists x(\delta(x)\wedge\varphi^i(x)\wedge t^i(x))]}{\exists x(\delta(x)\wedge\varphi^i(x))}\,\exists\text{E}^2$$

□

The next lemma will complete the simulation of ∀E and ∃I for non-variable terms, but requires a more elaborate proof.

**4.6.3 Lemma.** *Given an interpretation $i : T_1 \to T_2$, we have the following proofs for any $T_1$-formula $\chi(x)$ and non-variable term $t$:*

$$(\chi(t))^i \vdash^P_{T_2} \exists x(\delta(x) \wedge t^i(x) \wedge \chi^i(x))$$
$$\exists x(\delta(x) \wedge t^i(x) \wedge \chi^i(x)) \vdash^P_{T_2} (\chi(t))^i$$

*Proof.* We use induction on $\chi$.[10] 'IH' will refer to relevant applications of the induction hypothesis.

- **Base case.** $\chi(t)$ is a predicate $R(t)$. By the (adapted) definition of an interpretation, $(R(t))^i$ is equal to $\exists x(\delta(x) \wedge t^i(x) \wedge F(R)(x))$. Since $x$ is a variable, $F(R)(x)$ is exactly $R^i(x)$. So, by definition $(R(t))^i \leftrightarrow \exists x(\delta(x) \wedge t^i(x) \wedge R^i(x))$, which becomes a tautology instance.

- **Conjunction.** $\chi(t)$ is a formula $\varphi(t) \wedge \psi(t)$. We show both directions of the pure proof of $((\varphi(t))^i \wedge (\psi(t))^i) \leftrightarrow \exists x(\delta(x) \wedge t^i(x) \wedge \varphi^i(x) \wedge \psi^i(x))$.

  *Left-to-right direction.* We use first-order substitution in the proof. The proof is split up into three parts. Let [1] be $[\delta(w) \wedge t^i(w) \wedge \psi^i(w)]$, and let [2] be $[\delta(z) \wedge t^i(z) \wedge \varphi^i(z)]$.

$$\dfrac{\dfrac{\dfrac{\forall xy(\delta_{x,y} \to (t^i_{x,y} \to x =^i y))}{\delta_{z,w} \to (t^i_{z,w} \to z =^i w)}\,\forall\text{E} \quad \dfrac{\dfrac{[1]}{\delta(w)}\,\wedge\text{E} \quad \dfrac{[2]}{\delta(z)}\,\wedge\text{E}}{\delta_{z,w}\ (\approx)}\,\wedge\text{I}}{t^i_{z,w} \to z =^i w}\,\to\text{E} \quad \dfrac{\dfrac{[1]}{t^i(w)}\,\wedge\text{E} \quad \dfrac{[2]}{t^i(z)}\,\wedge\text{E}}{t^i_{z,w}}\,\wedge\text{I}}{z =^i w\ (*_1)}\,\to\text{E}$$

[10] Of the propositional connectives, for succinctness we only treat ∧ and →, which together also cover ¬ (which is defined as → ⊥) and ∨. Note that we do have ∨ in our language, but we choose to not present this case here because its classical definition is covered by the other connectives, and this paper restricts to the classical first-order natural deduction system.

$$\dfrac{\dfrac{*_1 \qquad \dfrac{\approx \qquad \dfrac{\forall xy(\delta_{x,y} \to (x =^i y \to \varphi^i(x) \to \varphi^i(y)))}{\delta_{z,w} \to (z =^i w \to \varphi^i(z) \to \varphi^i(w))}\forall\mathrm{E}}{z =^i w \to (\varphi^i(z) \to \varphi^i(w))}\to\mathrm{E} \qquad \dfrac{[2]}{\varphi^i(z)}\wedge\mathrm{E}}{\varphi^i(z) \to \varphi^i(w)}\to\mathrm{E} \qquad [1]}{\delta(w) \wedge t^i(w) \wedge \varphi^i(w) \wedge \psi^i(w)\ (*_2)}\wedge\mathrm{I}$$

$$\dfrac{\dfrac{\dfrac{*_2}{\exists x(\delta(x) \wedge t^i(x) \wedge \varphi^i(x) \wedge \psi^i(x))}\exists\mathrm{I} \qquad \dfrac{\dfrac{(\varphi(t))^i \wedge (\psi(t))^i}{(\varphi(t))^i}\wedge\mathrm{E}}{\exists x(\delta(x) \wedge t^i(x) \wedge \varphi^i(x))}\mathrm{IH}}{\exists x(\delta(x) \wedge t^i(x) \wedge \varphi^i(x) \wedge \psi^i(x))}\exists\mathrm{E}^1 \qquad \dfrac{\dfrac{(\varphi(t))^i \wedge (\psi(t))^i}{(\psi(t))^i}\wedge\mathrm{E}}{\exists x(\delta(x) \wedge t^i(x) \wedge \psi^i(x))}\mathrm{IH}}{\exists x(\delta(x) \wedge t^i(x) \wedge (\varphi^i(x) \wedge \psi^i(x)))}\exists\mathrm{E}^2$$

*Right-to-left direction.* Let $[1]$ be $[\delta(z) \wedge t^i(z) \wedge \varphi^i(z) \wedge \psi^i(z)]$. In the proof, the vertical dots indicate that this derivation is analogous to that of $(\varphi(t))^i$ in the left branch.

$$\dfrac{\dfrac{\dfrac{\dfrac{[1]}{\delta(z) \wedge t^i(z) \wedge \varphi^i(z)}\wedge\mathrm{E}}{\exists x(\delta(x) \wedge t^i(x) \wedge \varphi^i(x))}\exists\mathrm{I}}{(\varphi(t))^i}\mathrm{IH} \qquad \begin{array}{c}[1]\\ \vdots \\ (\psi(t))^i\end{array}}{(\varphi(t))^i \wedge (\psi(t))^i}\wedge\mathrm{I} \qquad \exists x(\delta(x) \wedge t^i(x) \wedge \varphi^i(x) \wedge \psi^i(x))}{(\varphi(t))^i \wedge (\psi(t))^i}\exists\mathrm{E}^1$$

- **Implication.** $\chi(t)$ is a formula $\varphi(t) \to \psi(t)$. We show both directions of the pure proof of $((\varphi(t))^i \to (\psi(t))^i) \leftrightarrow \exists x(\delta(x) \wedge t^i(x) \wedge (\varphi^i(x) \to \psi^i(x)))$.

  *Left-to-right direction.* The proof uses totality. Let $[1]$ be $[\varphi^i(w)]$, $[2]$ be $[\delta(w) \wedge t^i(w)]$ and $[3]$ be $[\delta(z) \wedge t^i(z) \wedge \psi^i(z)]$.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[2] \qquad [1]}{\delta(w) \wedge t^i(w) \wedge \varphi^i(w)}\wedge\mathrm{I}}{\exists x(\delta(x) \wedge t^i(x) \wedge \varphi^i(x))}\exists\mathrm{I}}{(\varphi(t))^i}\mathrm{IH} \qquad (\varphi(t))^i \to (\psi(t))^i}{(\psi(t))^i}\to\mathrm{E}}{\exists x(\delta(x) \wedge t^i(x) \wedge \psi^i(x))\ (*_1)}\mathrm{IH}$$

$$\dfrac{\dfrac{\dfrac{[2]}{t^i(w)} \qquad \dfrac{[3]}{t^i(z)}\wedge\mathrm{I}}{t^i_{w,z}} \qquad \dfrac{\dfrac{\dfrac{[2]}{\delta(w)} \qquad \dfrac{[3]}{\delta(z)}}{\delta_{w,z}\ (\approx)} \qquad \dfrac{\forall xy(\delta_{x,y} \to (t^i_{x,y} \to x =^i y))}{\delta_{w,z} \to (t^i_{w,z} \to w =^i z)}\forall\mathrm{E}}{t^i_{w,z} \to w =^i z}\to\mathrm{E}}{w =^i z\ (*_2)}$$

$$\dfrac{*_2 \qquad \dfrac{\dfrac{\forall xy(\delta_{x,y} \to (x =^i y \to (\psi^i(x) \to \psi^i(y))))}{\delta_{w,z} \to (w =^i z \to (\psi^i(z) \to \psi^i(w)))}\forall\mathrm{E} \qquad \approx}{w =^i z \to (\psi^i(z) \to \psi^i(w))}\to\mathrm{E}}{\psi^i(z) \to \psi^i(w)\ (*_3)}$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[3]}{\psi^i(z)}\wedge\mathrm{E}}{\psi^i(w)}{}_{*_3}\to\mathrm{E}}{\psi^i(w)}{}_{*_1}\exists\mathrm{E}^3}{\varphi^i(w)\to\psi^i(w)}\to\mathrm{I}^1 \quad [2]}{\delta(w)\wedge t^i(w)\wedge(\varphi^i(w)\to\psi^i(w))}\wedge\mathrm{I}}{\exists x((\delta(x)\wedge t^i(x))\wedge(\varphi^i(x)\to\psi^i(x)))}\exists\mathrm{I} \quad \exists x(\delta(x)\wedge t^i(x))}{\exists x((\delta(x)\wedge t^i(x))\wedge(\varphi^i(x)\to\psi^i(x)))}\exists\mathrm{E}^2$$

*Right-to-left direction.* Let [1] be $[\delta(w)\wedge t^i(w)\wedge\varphi^i(w)]$, [2] be $[\delta(z)\wedge t^i(z)\wedge(\varphi^i(z)\to\psi^i(z))]$, and [3] be $[(\varphi(t))^i]$.

$$\dfrac{\dfrac{\dfrac{\forall xy(\delta_{x,y}\to(t^i_{x,y}\to x=^i y))}{\delta_{w,z}\to(t^i_{w,z}\to w=^i z)}\forall\mathrm{E}^* \quad \dfrac{\dfrac{[1]}{\delta(w)} \quad \dfrac{[2]}{\delta(z)}}{\delta_{w,z}\ (\approx)}}{t^i_{w,z}\to w=^i z} \quad \dfrac{\dfrac{[1]}{t^i(w)} \quad \dfrac{[2]}{t^i(z)}}{t^i_{w,z}}\wedge\mathrm{I}}{w=^i z\ (*_1)}\to\mathrm{E}$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\forall xy(\delta_{x,y}\to x=^i y\to(\varphi^i(x)\to\varphi^i(y)))}{\delta_{w,z}\to(w=^i z\to(\varphi^i(w)\to\varphi^i(z)))}\to\mathrm{E} \quad \approx}{{}_{*_1}\ w=^i z\to(\varphi^i(w)\to\varphi^i(z))}\to\mathrm{I} \quad \dfrac{[1]}{\varphi^i(w)}\wedge\mathrm{E}}{\varphi^i(z)}\to\mathrm{E} \quad \dfrac{[2]}{\varphi^i(z)\to\psi^i(z)}\wedge\mathrm{E}}{\psi^i(z)\ (*_2)}\to\mathrm{E}$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[2]}{\delta(z)\wedge t^i(z)}\wedge\mathrm{E}}{\delta(z)\wedge t^i(z)\wedge\psi^i(z)}{}_{*_2}\wedge\mathrm{I}}{\exists x(\delta(x)\wedge t^i(x)\wedge\psi^i(x))}\exists\mathrm{I}}{(\psi(t))^i}\mathrm{IH} \quad \exists x(\delta(x)\wedge t^i(x)\wedge(\varphi(x))^i\to(\psi(x))^i)}{(\psi(t))^i}\exists\mathrm{E}^2 \quad \dfrac{[3]}{\exists x(\delta(x)\wedge t^i(x)\wedge\varphi^i(x))}\mathrm{IH}}{\dfrac{(\psi(t))^i}{(\varphi(t))^i\to(\psi(t))^i}\to\mathrm{I}^3}\exists\mathrm{E}^1$$

- **Universal quantifier.** $\chi(t)$ is a formula $\forall x\varphi(x,t)$. We show both directions of the pure proof of $(\forall x(\delta(x)\to(\varphi(x,t))^i)\leftrightarrow\exists x(\delta(x)\wedge t^i(x)\wedge\forall y(\delta(y)\to\varphi^i(y,x)))$. The induction hypothesis will hold for each instance $\varphi$.

  *Left-to-right direction:* The proof uses totality, uniqueness and first-order equality. [1] will stand for $[\delta(c)]$, [2] for $[\delta(h)\wedge t^i(h)\wedge\varphi^i(c,h)]$, [3] for $[\delta(e)\wedge t^i(e)\wedge\varphi^i(a,e)]$, and [4] for $[\delta(a)\wedge t^i(a)]$ (all the lowercase letters introduced here stand for variables).

$$\dfrac{\dfrac{\dfrac{[3]}{t^i(e)}\wedge\mathrm{E} \quad \dfrac{[2]}{t^i(h)}\wedge\mathrm{E}}{t^i_{e,h}} \quad \dfrac{\dfrac{\forall xy(\delta_{x,y}\to(t^i_{x,y}\to(x=^i y)))}{\delta_{e,h}\to(t^i_{e,h}\to(e=^i h))} \quad \dfrac{\dfrac{[3]}{\delta(e)} \quad \dfrac{[2]}{\delta(h)}}{\delta_{e,h}\ (\approx)}}{t^i_{e,h}\to(e=^i h)}}{e=^i h\ (*_0)}$$

$$\dfrac{*_0 \quad \dfrac{\dfrac{\forall xy(\delta_{x,y}\to(x=^i y\to(\varphi^i(c,x)\to\varphi^i(c,y))))}{\delta_{e,h}\to(e=^i h\to(\varphi^i(c,h)\to\varphi^i(c,e)))} \quad \approx}{e=^i h\to(\varphi^i(c,h)\to\varphi^i(c,e))}}{\varphi^i(c,h)\to\varphi^i(c,e)\ (*_1)}\to\mathrm{E}$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\forall x(\delta(x)\to(\varphi(x,t))^i)}{\delta(c)\to(\varphi(c,t))^i}\forall\mathrm{E} \quad [1]}{(\varphi(c,t))^i}\to\mathrm{E}}{\exists v(\delta(v)\wedge t^i(v)\wedge\varphi^i(c,v))}\mathrm{IH} \quad \dfrac{*_1 \quad \dfrac{[2]}{\varphi^i(c,h)}\wedge\mathrm{E}}{\varphi^i(c,e)}\to\mathrm{E}}{\varphi^i(c,e)}\exists\mathrm{E}^2}{\delta(c)\to\varphi^i(c,e)}\to\mathrm{I}^1}{\forall y(\delta(y)\to\varphi^i(y,e))\ (*_2)}\forall\mathrm{I}$$

$$\dfrac{\dfrac{\dfrac{\dfrac{*_2 \quad \dfrac{[3]}{\delta(e)\wedge t^i(e)}\wedge\mathrm{E}}{\delta(e)\wedge t^i(e)\wedge\forall y(\delta(y)\to\varphi^i(y,e))}\wedge\mathrm{I}}{\exists x(\delta(x)\wedge t^i(x)\wedge\forall y(\delta(y)\to\varphi^i(y,x)))}\exists\mathrm{I} \quad \dfrac{\dfrac{\dfrac{\forall x(\delta(x)\to(\varphi(x,t))^i)}{\delta(a)\to(\varphi(a,t))^i}\forall\mathrm{E} \quad \dfrac{[4]}{\delta(a)}\wedge\mathrm{E}}{(\varphi(a,t))^i}\to\mathrm{E}}{\exists b(\delta(b)\wedge t^i(b)\wedge\varphi^i(a,b))}\mathrm{IH}}{\exists x(\delta(x)\wedge t^i(x)\wedge\forall y(\delta(y)\to\varphi^i(y,x)))}\exists\mathrm{E}^3 \quad \exists x(\delta(x)\wedge t^i(x))}{\exists x(\delta(x)\wedge t^i(x)\wedge\forall y(\delta(y)\to\varphi^i(y,x)))}\exists\mathrm{E}^5$$

*Right-to-left direction:* Let [1] be $[\delta(b)\wedge t^i(b)\wedge\forall y(\delta(y)\to\psi^i(y,b))]$ and [2] be $[\delta(z)]$.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[1]}{\delta(b)\wedge t^i(b)}\wedge\mathrm{E} \quad \dfrac{\dfrac{\dfrac{[1]}{\forall y(\delta(y)\to\psi^i(y,b))}\wedge\mathrm{E}}{\delta(z)\to\psi^i(z,b)}\forall\mathrm{E} \quad [2]}{\psi^i(z,b)}\to\mathrm{E}}{\delta(b)\wedge t^i(b)\wedge\psi^i(z,b)}\wedge\mathrm{I}}{\exists x(\delta(x)\wedge t^i(x)\wedge\psi^i(z,x))}\exists\mathrm{I}}{(\psi(z,t))^i}\mathrm{IH}}{\delta(z)\to(\psi(z,t))^i}\to\mathrm{I}^2}{\forall y(\delta(y)\to(\psi(y,t))^i)}\forall\mathrm{I} \quad \exists x(\delta(x)\wedge t^i(x)\wedge\forall y(\delta(y)\to\psi^i(y,x)))}{\forall y(\delta(y)\to(\psi(y,t))^i)}\exists\mathrm{E}^1$$

- **Existential quantifier.** $\chi(t)$ is a formula $\exists y\varphi(y,t)$. We show both directions of the pure proof of $\exists x(\delta(x)\wedge(\varphi(x,t))^i)\leftrightarrow\exists y(\delta(y)\wedge t^i(y)\wedge\exists x(\delta(x)\wedge(\varphi(x,y))^i)$.

  *Left-to-right direction:* Let [1] be $[\delta(a)\wedge(\varphi(a,t))^i]$, and [2] be $[\delta(b)\wedge t^i(b)\wedge\varphi^i(a,b)]$.

$$\dfrac{\dfrac{\dfrac{[1]}{(\varphi(a,t))^i}\,\wedge\mathrm{E}}{\exists x(\delta(x)\wedge t^i(x)\wedge\varphi^i(a,x))}\,\mathrm{IH} \qquad \dfrac{\dfrac{\dfrac{\dfrac{[1]}{\delta(a)}\,\wedge\mathrm{E} \quad \dfrac{[2]}{\varphi^i(a,b)}\,\wedge\mathrm{E}}{\delta(a)\wedge\varphi^i(a,b)}\,\wedge\mathrm{I}}{\exists y(\delta(y)\wedge\varphi^i(y,b))}\,\exists\mathrm{I} \qquad \exists x(\delta(x)\wedge(\varphi(x,t))^i)}{\exists y(\delta(y)\wedge\varphi^i(y,b))}\,\exists\mathrm{E}^1 \quad \dfrac{[2]}{\delta(b)\wedge t^i(b)}\,\wedge\mathrm{E}}{\dfrac{\delta(b)\wedge t^i(b)\wedge\exists y(\delta(y)\wedge\varphi^i(y,b))}{\exists x(\delta(x)\wedge t^i(x)\wedge\exists y(\delta(y)\wedge\varphi^i(y,x)))}\,\exists\mathrm{I}}\,\wedge\mathrm{I}}{\exists x(\delta(x)\wedge t^i(x)\wedge\exists y(\delta(y)\wedge\varphi^i(y,x))))}\,\exists\mathrm{E}^2$$

*Right-to-left direction*. Let [1] be $[\delta(c)\wedge\varphi^i(c,a)]$, and [2] be $[\delta(a)\wedge t^i(a)\wedge\exists y(\delta(y)\wedge\varphi^i(y,a))]$.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[1]}{\varphi^i(c,a)}\,\wedge\mathrm{E} \quad \dfrac{[2]}{\delta(a)\wedge t^i(a)}\,\wedge\mathrm{E}}{\delta(a)\wedge t^i(a)\wedge\varphi^i(c,a)}\,\wedge\mathrm{I}}{\exists x(\delta(x)\wedge t^i(x)\wedge\varphi^i(c,x))}\,\exists\mathrm{I}}{(\varphi(c,t))^i}\,\mathrm{IH} \quad \dfrac{[1]}{\delta(c)}\,\wedge\mathrm{E}}{\dfrac{\delta(c)\wedge(\varphi(c,t))^i}{\exists y(\delta(y)\wedge(\varphi(y,t))^i)}\,\exists\mathrm{I}}\,\wedge\mathrm{I} \qquad \dfrac{[2]}{\exists y(\delta(y)\wedge\varphi^i(y,a))}\,\wedge\mathrm{E}}{\dfrac{\exists y(\delta(y)\wedge(\varphi(y,t))^i) \qquad \exists x(\delta(x)\wedge t^i(x)\wedge\exists y(\delta(y)\wedge(\varphi(y,x))^i))}{\exists y(\delta(y)\wedge(\varphi(y,t))^i)}\,\exists\mathrm{E}^2}\,\exists\mathrm{E}^1$$

□

Now we are ready to give the full proof of the theorem, which we repeat here.

**4.6.4 Theorem.** *Let $i : \mathsf{T}_1 \to \mathsf{T}_2$ be an interpretation. Let $\mathcal{D}$ be the proof $\Gamma \vdash_{\mathsf{T}_1} \varphi$ in the classical first-order natural deduction calculus. Then $\Gamma^i, \delta_{\lambda_\mathcal{D}} \vdash^P_{\mathsf{T}_2} \varphi^i$ by simulation.*

*Proof.* We prove this by induction on the length of the derivation of $\Gamma \vdash_{\mathsf{T}_1} \varphi$.

- **Base case.** $\varphi$ is a $\mathsf{T}_1$-axiom or -assumption. By definition of an interpretation, there exists a (trivially secondarily ontologically pure) proof $\delta_\varphi \vdash^P_{\mathsf{T}_2} \varphi^i$.

- **Case $\wedge$I.** Suppose $\varphi$ equals $\chi \wedge \psi$ and the last step of $\mathcal{D}$ was $\dfrac{\chi \quad \psi}{\chi\wedge\psi}\,\wedge I$. This means there are subderivations $\mathcal{D}_1$ referring to $\Gamma \vdash_{\mathsf{T}_1} \chi$, and $\mathcal{D}_2$ referring to $\Gamma \vdash_{\mathsf{T}_1} \psi$.[11] Then by the induction hypothesis, we have pure simulations $\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}} \vdash^P_{\mathsf{T}_2} \chi^i$, and $\Gamma^i, \delta_{\lambda_{\mathcal{D}_2}} \vdash^P_{\mathsf{T}_2} \psi^i$. Then we obtain the following pure simulation of the whole proof $\mathcal{D}$:

[11] $\mathcal{D}_1$ and $\mathcal{D}_2$ may only use a subset of $\Gamma$, but for easy notation we will use the entire set $\Gamma$, which will only require a conjunction elimination in the worst case, and we can easily transform a pure simulation using a subset of $\Gamma^i$ into one that starts from $\Gamma^i$ as a whole.

$$\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}}}\wedge\mathrm{E} \qquad \dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\Gamma^i, \delta_{\lambda_{\mathcal{D}_2}}}\wedge\mathrm{E}}{\dfrac{\begin{array}{cc}\vdots & \vdots \\ \chi^i & \psi^i\end{array}}{\chi^i \wedge \psi^i}\wedge\mathrm{I}}$$

This works by definition of the interpretation, as $\chi^i \wedge \psi^i = (\chi \wedge \psi)^i$. The case $\wedge$E works similarly, which we omit.

- **Case $\vee$I.** Suppose $\varphi$ equals $\chi \vee \psi$ and the last step of $\mathcal{D}$ was $\dfrac{\chi}{\chi \vee \psi}\vee I$ . This means there is a subderivation $\mathcal{D}_1$ referring to $\Gamma \vdash_{\mathsf{T}_1} \chi$. By the induction hypothesis, we have a pure simulation $\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}} \vdash^P_{\mathsf{T}_2} \chi^i$. Then we obtain the following pure simulation of the whole proof $\mathcal{D}$:

$$\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}}}\wedge\mathrm{E}}{\dfrac{\vdots}{\chi^i \vee \psi^i}\vee I^i}$$

This works because of the interpretation, as $\chi^i \vee \psi^i = (\chi \vee \psi)^i$, and because certainly the introduction of $\psi^i$ is valid according to the restricted rule $\vee\mathrm{I}^i$.

- **Case $\vee$E.** Suppose $\varphi$ was obtained by the $\vee$E-rule, i.e. $\dfrac{\chi \vee \psi \qquad \begin{array}{c}[\chi] \\ \vdots \\ \varphi\end{array} \qquad \begin{array}{c}[\psi] \\ \vdots \\ \varphi\end{array}}{\varphi}\vee E$ . This means there are subderivations $\mathcal{D}_1$ referring to $\Gamma \vdash_{\mathsf{T}_1} \chi \vee \psi$, $\mathcal{D}_2$ referring to $\chi \vdash_{\mathsf{T}_1} \varphi$ and $\mathcal{D}_3$ referring to $\psi \vdash_{\mathsf{T}_1} \varphi$. By the induction hypothesis, we have pure simulations $\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}} \vdash^P_{\mathsf{T}_2} (\chi \vee \psi)^i$, $\varphi^i, \delta_{\lambda_{\mathcal{D}_2}} \vdash^P_{\mathsf{T}_2} \varphi^i$ and $\psi^i, \delta_{\lambda_{\mathcal{D}_3}} \vdash^P_{\mathsf{T}_2} \varphi^i$. Then we obtain the following pure simulation of the whole proof $\mathcal{D}$.

$$\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}}}\wedge\mathrm{E} \qquad \dfrac{[\chi^i] \qquad \dfrac{[\delta_{\lambda_{\mathcal{D}}}]}{\delta_{\lambda_{\mathcal{D}_2}}}\wedge\mathrm{E}}{\delta_{\lambda_{\mathcal{D}_2}}, \chi^i}\wedge\mathrm{I} \qquad \dfrac{[\psi^i] \qquad \dfrac{[\delta_{\lambda_{\mathcal{D}}}]}{\delta_{\lambda_{\mathcal{D}_3}}}\wedge\mathrm{E}}{\delta_{\lambda_{\mathcal{D}_3}}, \psi^i}\wedge\mathrm{I}}{\dfrac{\begin{array}{ccc}\vdots & \vdots & \vdots \\ (\chi \vee \psi)^i & \varphi^i & \varphi^i\end{array}}{\varphi^i}\vee E}$$

Again, this works because we may see $(\chi \vee \psi)^i$ in the premise of the rule as $\chi^i \vee \psi^i$, so that it is ready for a disjunction elimination rule to be applied to it. The cases of $\rightarrow$I and $\rightarrow$E work similarly, which we omit.

- **Case** $\forall$**I.** Suppose $\varphi$ equals $\forall x\chi$ and the last step of $\mathcal{D}$ was $\dfrac{\chi[x\backslash y]}{\forall x\chi}\,\forall I$ . This means there is a subderivation $\mathcal{D}_1$ referring to $\Gamma \vdash_{\mathsf{T}_1} \chi[x\backslash y]$. By the induction hypothesis, we have the pure simulation $\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}} \vdash^P_{\mathsf{T}_2} \chi^i(y)$. Then we obtain the following pure simulation of the whole proof $\mathcal{D}$. In this case, we can see that $\delta_{\lambda_{\mathcal{D}}} = \delta_{\lambda_{\mathcal{D}_1}}$.

$$\dfrac{\dfrac{\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\vdots}}{\chi^i(y)} \qquad \dfrac{}{[\delta(y)]^1}\,\wedge\mathrm{I}}{\chi^i(y) \wedge \delta(y)}\,\wedge\mathrm{E}}{\delta(y) \rightarrow \chi^i(y)}\,\rightarrow\mathrm{I}^1}{\forall x(\delta(x) \rightarrow \chi^i(x))}\,\forall\mathrm{I}$$

- **Case** $\exists$**E.** Suppose $\varphi$ was obtained by the $\exists$E-rule, i.e. $\dfrac{\exists x\chi(x) \qquad \begin{matrix}[\chi(y)]\\ \vdots \\ \varphi\end{matrix}}{\varphi}\,\exists E$ . This means there are subderivations $\mathcal{D}_1$ referring to $\Gamma \vdash_{\mathsf{T}_1} \exists x\chi(x)$ and $\mathcal{D}_2$ referring to $\chi(y) \vdash_{\mathsf{T}_1} \varphi$. By the induction hypothesis, there exist pure simulations $\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}} \vdash^P_{\mathsf{T}_2} \exists x(\delta(x) \wedge \chi^i(x))$ and $\chi^i(y), \delta_{\lambda_{\mathcal{D}_2}} \vdash^P_{\mathsf{T}_2} \varphi^i$. Then in $\mathsf{T}_2$, there exists the following pure simulation of the whole proof $\mathcal{D}$.

$$\dfrac{\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}}}\,\wedge\mathrm{E} \qquad \dfrac{\dfrac{[\delta_{\lambda_{\mathcal{D}}}]}{\delta_{\lambda_{\mathcal{D}_2}}}\,\wedge\mathrm{E} \qquad \dfrac{[\delta(y) \wedge \chi^i(y)]^1}{\chi^i(y)}\,\wedge\mathrm{E}}{\delta_{\lambda_{\mathcal{D}_2}}, \chi^i(y)}\,\wedge\mathrm{I}}{\begin{matrix}\vdots \\ \varphi^i\end{matrix}}}{\exists x(\delta(x) \wedge \chi^i(x)) \qquad \begin{matrix}\vdots \\ \varphi^i\end{matrix}}\,\exists\mathrm{E}^1}{\varphi^i}$$

- **Case** $\forall$**E.** Suppose $\varphi$ equals $\chi(t)$ and the last step of $\mathcal{D}$ was $\dfrac{\forall x\chi(x)}{\chi(t)}\,\forall\mathrm{E}$ . This means there is a subderivation $\mathcal{D}_1$ referring to $\Gamma \vdash_{\mathsf{T}_1} \forall x\chi(x)$. By the induction hypothesis, we have a pure simulation $\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}} \vdash^P_{\mathsf{T}_2} \forall x(\delta(x) \rightarrow \chi^i(x))$. Now we distinguish between two cases. For *variable* terms (so that $(\chi(t))^i = \chi^i(t)$), the simulation of this rule will consist of:

$$\dfrac{\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}}}}{\vdots}}{\dfrac{\dfrac{\forall x(\delta(x) \to \chi^i(x))}{\delta(t) \to \chi^i(t)}\,\forall\mathrm{E} \qquad \dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\delta(t)}\,\wedge\mathrm{E}}{\chi^i(t)}\,\to\mathrm{E}}$$

For *non-variable* terms, we can see that $\delta_{\lambda_{\mathcal{D}}} = \delta_{\lambda_{\mathcal{D}_1}}$. The simulation of the rule consists of the following (double lines indicate an application of the relevant lemma):

$$\dfrac{\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\vdots}}{\forall x(\delta(x) \to \chi^i(x))}}{\exists x(\delta(x) \wedge t^i(x) \wedge \chi^i(x))}\,\text{Lemma 4.6.2}}{(\chi(t))^i}\,\text{Lemma 4.6.3}$$

- **Case** $\exists$**I**. Suppose $\varphi$ equals $\exists x\chi(x)$ and the last step of $\mathcal{D}$ was $\dfrac{\chi(t)}{\exists x\chi(x)}\,\exists\mathrm{I}$. This means there is a subderivation $\mathcal{D}_1$ referring to $\Gamma \vdash_{\mathsf{T}_1} \chi(t)$. By the induction hypothesis, we have a pure simulation $\Gamma^i, \delta_{\lambda_{\mathcal{D}_1}} \vdash^P_{\mathsf{T}_2} (\chi(t))^i$. In this case, again, we can see that $\delta_{\lambda_{\mathcal{D}}} = \delta_{\lambda_{\mathcal{D}_1}}$. We again distinguish between two cases. For *variable* terms, the simulation of this rule will consist of:

$$\dfrac{\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\vdots}}{\chi^i(t)} \qquad \dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\delta(t)}\,\wedge\mathrm{E}}{\delta(t) \wedge \chi^i(t)}}{\exists x(\delta(x) \wedge \chi^i(x))}$$

For *non-variable* terms, the simulation of the rule consists of:

$$\dfrac{\dfrac{\dfrac{[\Gamma^i, \delta_{\lambda_{\mathcal{D}}}]}{\vdots}}{(\chi(t))^i}}{\exists x(\delta(x) \wedge t^i(x) \wedge \chi^i(x))}\,\text{Lemma 4.6.3}}{(\exists x\chi(x))^i}\,\text{Lemma 4.6.2}$$

□

# 5

# An ideal of proof systems

## *Preliminaries to the analysis of semantic pollution*

This chapter will provide the conceptual as well as technical background for the analysis of a particular ideal of proof systems in Chapter 6. This means that we now transition away from ideals of informal (mathematical) proofs, and we focus on properties of the design of a proof system itself. In particular, we will zoom in on proof systems for modal logic in order to introduce and formalize the ideal of 'syntactic purity', or its more well-known counterpart, 'semantic pollution'. As a start to doing so, in this chapter we provide the background for this phenomenon, and we familiarize ourselves with the variety of formal systems that characterize modal logic.

First, we briefly consider the difficulties encountered in designing proof systems for modal logic in Section 5.1. In Section 5.2, we will then introduce the grammar and model-theoretic interpretation of the most common proof systems for modal logic. Instead of providing the full definition of a proof system in terms of grammar, axioms and inference rules, we will focus mainly on their grammar (that we say generates a *proof-theoretic language*). After that, Section 5.3 will describe the usual notions of equivalence for Kripke models (the most common type of models for modal logic), as well as for Kripke models extended by an assignment function. The introduction of these languages and model equivalences will involve some of the formal notation that we are going to use in the next chapter. We end the chapter in Section 5.4 by considering several properties that intuitively relate to semantic pollution, that we nevertheless reject; and so we are ready to embark on the full proposal for semantic pollution in Chapter 6. The work in this chapter corresponds to parts of the publication (Martinot, 2022), and part of the submitted work (Martinot, 2024b).

## 5.1 Introduction

Modal logic is the logic obtained from propositional logic by adding modalities to the language (usually denoted by $\Box$ and $\Diamond$), which are operators used to qualify the truth of a statement. The origin of studying modal logic is often attributed to Lewis (1918), who introduced the first modal operators with the aim of solving paradoxes of material implication. Nowadays, a plentitude of modal operators have been introduced and analyzed, all of which express new 'modes of truth'. Common modalities are *alethic* modalities (that stand for necessity and possibility), *epistemic* modalities (representing knowledge and belief), *temporal* modalities (relativizing truth to the past or future), and *provability* modalities (representing provability in arithmetical theories). More information on the history of modal logic may be found for instance in (Blackburn et al., 2001, §1.7).

While modal logics have wide applications in computer science, AI, game theory and philosophical logic, the consensus is that 'Gentzen's proof-theoretical methods' have not flourished particularly well within the domain of modal logic (see e.g. (Bull and Segerberg, 1984)). That is, the presentation of modal logics as 'ordinary' natural deduction and sequent calculi has proven troublesome for technical reasons, as well as philosophical ones. In particular, natural deduction calculi for modal logics have proven to be difficult to normalize, and sequent calculi for modal logics do not easily obtain the property of analyticity (that induces the subformula property) (Negri, 2011). Sambin and Valentini (1982) for instance illustrate the latter by attempting to find a satisfactory sequent calculus for the logic K4.

The basic modal logic K is obtained by taking the axioms of CPL (classical propositional logic) and the rule of Modus Ponens, and by adding the axiom $K$ ($\Box(A \to B) \to (\Box A \to \Box B)$) as well as the rule of *necessitation*: from $A$ infer $\Box A$. The logic K4 is in turn obtained from K by adding the axiom for transitivity ($\Box p \to \Box\Box p$), which Sambin and Valentini propose to add as a rule as follows.

$$\frac{\Gamma \Rightarrow \Box A, \Delta}{\Gamma \Rightarrow \Box\Box A, \Delta}$$

While this rule is perfectly adequate for establishing soundness and completeness results with respect to K4, it does not allow cut elimination.

Certain philosophically desirable properties of inference rules have also been found challenging to establish in the modal setting, such as the want for rules that are *separable* (Poggiolesi and Restall, 2012), *symmetric* and *explicit* (Wansing, 1994) (see also Chapter 1). For instance, one can introduce a rule for necessity in a proof system by turning the axiom $K$ into a rule (see also (Pimentel, 2018)). A sequent calculus version of this rule can be made as follows:

$$\frac{A_1, ..., A_n \Rightarrow B}{\Box A_1, ..., \Box A_n \Rightarrow \Box B}$$

But such a rule introduces the modality simultaneously on the left and on the right of the sequent arrow, and so introduces multiple of the same connective at once (violating the requirement for explicitness, among others). Instead, proof systems that introduce separate introduction and elimination rules provide better candidates for philosophically meaning-conferring rules, yet the standard attempts are not without problems, either. For instance, a standard modal (natural deduction) rule in Curry (1950)'s T-system for formal deducibility is the following:

$$\frac{A}{\Diamond A}\,\Diamond\mathsf{I}$$

As discussed in (Read, 2015), an elimination rule that is suitably in *harmony* with the introduction rule is in fact hard to find. For the interested reader, more details on the mentioned desirable properties of proof systems for modal logic are given by the references provided above (as well as, for instance, (Poggiolesi, 2010)).

In order to obtain modal inference rules with nicer properties, enriching the language of proof calculi has been a widespread solution. Below, we will introduce the most common such enriched languages, and the model-theoretic interpretation of their formulas. We will not specifically introduce the axioms and inference rules of proof calculi that use these languages, because they are not necessary for our characterization of semantic pollution in Chapter 6, and they will take up too much space here. However, references where full descriptions of the proof systems are given will be provided.

## 5.2 Syntax and semantics of proof systems for modal logic

We will introduce formal languages by their grammar, displaying the form of atomic formulas, and that of complex formulas. A language $\mathsf{L}$ is then equal to the set of well-formed formulas by this grammar.

**5.2.1 Definition** (**Formula type**)**.** Suppose $C$ is an $n$-ary operator in $\mathsf{L}$. Then for any subset (that we may call a '*context language*') $\mathsf{CL} \subseteq \mathsf{L}$, its *formula type* is denoted by:

$$C(\mathsf{CL}) := \{C(A_1, ..., A_n) \mid A_i \in \mathsf{CL} \text{ for } 1 \leq i \leq n\}$$

In case $C$ is atomic, its formula type is just the set of all its available occurrences.

We will use this definition later on more. For now, we can start by introducing the modal language, which we will from now on refer to as the *object language* $\mathsf{L}$. Let $\mathsf{Prop} = \{p, q, r, ...\}$ be a set of countably many propositional variables. Then $\mathsf{L}$ will refer to the (classical) basic modal language and contains the well-formed formulas of the grammar:

$$A ::= p \mid \bot \mid A \wedge A \mid A \vee A \mid A \to A \mid \Box A \qquad \text{(where } p \in \mathsf{Prop}\text{)}$$

Given a formula $A$, we abbreviate $\neg A$ as $A \to \bot$, $\top$ as $\neg\bot$ and $\Diamond A$ as $\neg\Box\neg A$. Thus, for each modal operator, its natural context language is $\mathsf{L}$ itself.

As the model theory for $\mathsf{L}$, we take the usual Kripke semantics, as specified e.g. in (Blackburn et al., 2001).

**5.2.2 Definition.** A *classical modal Kripke frame* $\mathsf{F}$ is a pair $(W, R)$, consisting of a set of states $W$ and a binary accessibility relation $R \subseteq W \times W$. A *classical modal Kripke model* $\mathsf{M}$ is a triple $(W, R, V)$, consisting of a Kripke frame $(W, R)$ together with a propositional valuation function $V : \mathsf{Prop} \to \mathcal{P}(W)$.

$\mathsf{L}$-formulas receive classical truth conditions relative to a pointed model, as follows.

**5.2.3 Definition.** Let $\mathsf{M} = (W, R, V)$ be a Kripke model, and let $w \in W$. The truth conditions of $\mathsf{L}$-formulas are then defined inductively as follows:

$\mathsf{M}, w \vDash p$ iff $w \in V(p)$

$\mathsf{M}, w \nvDash \bot$

$\mathsf{M}, w \vDash A \wedge B$ iff $\mathsf{M}, w \vDash A$ and $\mathsf{M}, w \vDash B$

$\mathsf{M}, w \vDash A \vee B$ iff $\mathsf{M}, w \vDash A$ or $\mathsf{M}, w \vDash B$

$\mathsf{M}, w \vDash A \to B$ iff $\mathsf{M}, w \vDash A$ implies $\mathsf{M}, w \vDash B$

$\mathsf{M}, w \vDash \Box A$ iff for all $v$ such that $wRv$, $\mathsf{M}, v \vDash A$

A *proof-theoretic language* $\mathsf{PL}$, generated by the grammar of a proof system, can then extend $\mathsf{L}$ by any new syntax. We will say that a proof system, and in particular its collection of axioms and inference rules, are 'based in' a proof-theoretic language. Among the proof-theoretic languages extending $\mathsf{L}$, we can distinguish between 'internal' and 'external' proof systems. As mentioned by Lyon et al. (2023), "the proof-theoretic community lacks consensus on how ['internal' and 'external' calculus] should be precisely defined". A common interpretation says that internal calculi allow every sequent to be interpretable as a formula in the logic, whereas external calculi do not. Lyon et al. (2023) makes the notion of a formula interpretation more precise: "what is meant [...] is a *translation* $\tau$ that maps every sequent to a (i) 'structurally similar' and (ii) 'logically equivalent' formula in the language of the logic". Then $\tau$ can be said to be a *homomorphism* preserving the structure of the sequent, and such that "a sequent is satisfied on a model of the underlying logic *iff* its output is". Lyon notes two downsides of this conceptualization: first, internality or externality is then dependent on the specific subset of sequents based in a particular language, which seems counterintuitive. Second, it appears that establishing that a language is external, i.e. confirming the non-existence of a suitable translation to the logic, can be a rather difficult task.

We will make sense of the distinction between internal and external calculi by making two changes to Lyon's conception. First, we will consider translations from the proof-theoretic language to the logic not to act on entire *sequents*, but on individual grammar categories (i.e., individual formula types) within the proof-theoretic language. The main motivation for this is that we aim to consider the nature of the separate operators that are added to the modal language, instead of considering sequents as one 'undividable block'. This slightly changes the obstacle of the property of internal/external syntax being dependent on the language subset you look at: now it is dependent on the operator itself, as well as its context language. Furthermore, we prefer to precisify the notion of internal and external calculi independently from the particular semantics, and so we specify it with respect to the logical entailment relation. For that, we consider $\vdash$ to refer to the derivability relation of a proof system based in a language $\mathsf{PL}$, and $\vdash_{\mathrm{LE}}$ to refer to the logical entailment relation of $\mathsf{L}$.[1]

**5.2.4 Definition (Formula interpretation).** Let $\mathsf{PL}$ be a proof-theoretic language extending $\mathsf{L}$, and consider a proof system based in $\mathsf{PL}$. Then $\mathsf{PL}$ will have a *formula interpretation* into $\mathsf{L}$, if there exists a translation function $t : \mathsf{PL} \to \mathsf{L}$ from proof-theoretic symbols $A$ to formulas in the logic[2], such that $B \vdash A$ implies $t(B) \vdash_{\mathrm{LE}} t(A)$. We will use 'having a formula interpretation in the logic' synonymously with 'being translatable to the logic'.

**5.2.5 Definition (Internal and external proof system).** A proof system is *internal* with respect to $\mathsf{L}$, if each element of its proof-theoretic language $\mathsf{PL}$ has a formula interpretation in $\mathsf{L}$. A proof system is *external* with respect to $\mathsf{L}$, if there is an expression of $\mathsf{PL}$ that does not have a formula interpretation in $\mathsf{L}$.

While we maintain the above definitions, we remain aware that the literature contains variants of them with possibly different properties. Our choice of presenting these notions will be compatible with the characterization of semantic pollution in the next chapter.

We now present several proof-theoretic languages for modal logic, as well as the truth conditions of the formulas they introduce in terms of Kripke semantics. The last three languages that we introduce will receive two-letter abbreviations, as they will serve as main case studies in Chapter 6 on semantic pollution. First, we mention some well-known structural generalizations of the sequent calculus. Consider the *hypersequent calculus* (after (Avron, 1987; Pottinger, 1983)), that generalizes the notion of a sequent by working with finite lists of sequents. A hypersequent looks as follows:

[1] Here, logical entailment refers to the consequence relation of the logic, satisfying appropriate versions of properties such as commonly, among others, reflexivity and transitivity.

[2] $t$ should have certain suitable properties like compositionality, to avoid triviality of the translation.

$$\Gamma_1 \Rightarrow \Delta_1 \mid ... \mid \Gamma_n \Rightarrow \Delta_n$$[3]

A hypersequent is interpreted commonly as a disjunction of sequents, so that the above expression is equivalent to $(\bigwedge \Gamma_1 \rightarrow \bigvee \Delta_1) \vee ... \vee (\bigwedge \Gamma_n \rightarrow \bigvee \Delta_n)$. Hypersequents are "analytic, but the presence of internal and external structural rules, usually non-eliminable, make them less suitable for the purposes of automated deduction" (Negri, 2011). A further generalization of the sequent calculus induces a tree structure into the syntax of a sequent. This is achieved by the nested sequent calculus (Brünnler, 2010) and the tree-hypersequent calculus (Poggiolesi, 2009), which are notational variants of each other (and which are predated by Bull (1992); Kashima (1994)). Here, any structure

$$A_1, ..., A_n, [\Delta_1], ..., [\Delta_m]$$

is a *nested sequent*, where $A_1, ..., A_n$ are modal formulas, and $\Delta_1, ...\Delta_m$ are nested sequents (a nested sequent is then a multiset of formulas and boxed sequents). Similarly, any structure

$$\Gamma / G_1; ...; G_n$$

is a tree-hypersequent, where $\Gamma$ is a set of modal formulas, and $G_1, ..., G_n$ are tree-hypersequents. The formula interpretation $()^{\mathsf{FI}}$ of nested and tree-hypersequents is then the same:

$$(A_1, ..., A_n, [\Delta_1], ..., [\Delta_m])^{\mathsf{FI}} := A_1 \vee ... \vee A_n \vee \Box\Delta_1^{\mathsf{FI}} \vee ... \vee \Delta_m^{\mathsf{FI}}$$
$$(\Gamma / G_1; ...; G_n)^{\mathsf{FI}} := \bigvee \Gamma \vee \Box G_1^{\mathsf{FI}} \vee ... \vee \Box G_n^{\mathsf{FI}}$$

Furthermore, consider now two languages that introduce an additional set of variables to the syntax of the proof system. The first is the *labeled language* (LL), and the second is the *hybrid language* (HL). The introduced variables are called 'labels' in labeled calculi and 'nominals' in hybrid logic. While labels are intuitively understood as naming states in a Kripke model, nominals were motivated by the want to formalize natural language sentences referring to specific time points or individuals. We will treat labels and nominals uniformly throughout this thesis, and call them *name variables* for both languages, given by a set $\mathsf{Var} = \{a, b, c, ..., x, y, z, ...\}$.[4] Formulas including name variables will receive truth conditions in terms of Kripke models extended by an assignment function $\tau : \mathsf{Var} \rightarrow W$.

The notation $\mathsf{M}$ will always stand for a regular Kripke model $(W, R, V)$ as in Definition 5.2.3. A *model for a proof-theoretic language* $\mathsf{PM}$ will either equal an extended model $(\mathsf{M}, \tau)$ (in case of LL and HL), or a regular Kripke model $\mathsf{M}$ (in case

[3] This notation should not be confused with the notation for grammar categories.

[4] In the literature, it is common for nominals to use early-occurring alphabet letters $a, b, c, ...$, and for labels to use late-occurring alphabet letters $x, y, z, ....$ In accordance with this custom, our examples for HL will commonly use letters $a, b, c, ...$, and our examples for LL will commonly use letters $x, y, z, ....$

of other proof-theoretic languages). Where it is relevant to know if we are talking about $\mathsf{M}$ or a pair $(\mathsf{M}, \tau)$, we will always use the specific notation. Similarly, a *frame for a proof-theoretic language* $\mathsf{PF}$ will either stand for a regular Kripke frame $\mathsf{F} = (W, R)$ or for an extended frame $(\mathsf{F}, \tau)$.

Now consider the proof-theoretic languages. The language $\mathsf{LL}$ uses the name variables in $\mathsf{Var}$ (as just defined above) to accompany object language formulas and to form new atomic formulas $xRy$. Thus, $\mathsf{LL}$ extends $\mathsf{L}$ by the following formulas, based on the labeled calculus as in (Negri, 2005) (note that labeled calculi date back to Simpson (1994); Kanger (1957)):

$$A ::= B \mid x : B \mid xRy \qquad \text{(where } B \in \mathsf{L} \text{ and } x, y \in \mathsf{Var}\text{)}$$

The truth conditions of the new proof-theoretic formulas are then defined as follows, for a Kripke model extended by assignment function $\tau$.

$$\mathsf{M}, \tau, w \vDash x : A \text{ iff } \mathsf{M}, \tau(x) \vDash A$$
$$\mathsf{M}, \tau, w \vDash xRy \text{ iff } \tau(x) R \tau(y)$$

Although we only focus on proof systems relative to Kripke models here, note that the labeled calculus can be seen to inspire a general approach to designing proof systems based on a certain semantics. See (Negri, 2016) for the general approach and for an elaboration of a proof system based on neighborhood semantics.

Next, the language $\mathsf{HL}$ uses the name variables in $\mathsf{Var}$ to introduce various new operators. Some of them are closely related to the labeled calculus, and thus $\mathsf{HL}$ provides a good comparison for semantic pollution in the next chapter.[5] $\mathsf{HL}$ extends the modal language by the following grammar, based on propositional hybrid logic as in (Braüner, 2010), which is a logical extension of the basic modal language.

$$A ::= p \mid a \mid \bot \mid A \wedge A \mid A \to A \mid \Box A \mid @_a A \mid \forall a A \mid {\downarrow} a A$$
$$\text{(where } p \in \mathsf{Prop}, a \in \mathsf{Var}\text{)}$$

The truth conditions of the hybrid formulas are as follows, again for a Kripke model extended by an assignment function $\tau$. Let $\tau_a$ be the assignment function that agrees with $\tau$ on every nominal assignment, except possibly on $a$. Specifically, let $\tau_{[a \mapsto w]}$ be the assignment function that agrees with $\tau$ on every nominal

[5] Technically, as $\mathsf{HL}$ is primarily an object language, and secondarily a proof-theoretic language, results of semantic pollution will apply to this language in both roles. We thus take semantic pollution then generally to apply to any language *extending the basic modal language*, not just to proof-theoretic languages, although of course the use of languages in proof systems is what concerns the debate on semantic pollution most, and what we focus on here.

assignment, except possibly on $a$, which is sent to world $w$.

$$\mathsf{M}, \tau, w \vDash a \text{ iff } \tau(a) = w$$
$$\mathsf{M}, \tau, w \vDash @_a A \text{ iff } \mathsf{M}, \tau, \tau(a) \vDash A$$
$$\mathsf{M}, \tau, w \vDash \forall a A \text{ iff for any } \tau_a, \mathsf{M}, \tau_a, w \vDash A$$
$$\mathsf{M}, \tau, w \vDash \downarrow a A \text{ iff } \mathsf{M}, \tau_{[a \mapsto w]}, w \vDash A$$

Finally, we present the *display language* (DL). DL extends L by forming the following grammar, based on the display calculus as in (Wansing, 1994) (an extension of (Belnap, 1982)):

$$A ::= B \mid \mathbf{I} \mid (A \circ A) \mid *A \mid \bullet A \qquad \text{(where } B \in \mathsf{L}\text{)}$$

That is, the display calculus adds new structural connectives **I**, $\circ$, $*$, and $\bullet$ to the modal language. The resulting structures have two different translations into tense logic, depending on their position (antecedent or consequent) in a sequent. The operators **I**, $\circ$ and $*$ have intended translations in terms of conjunction, disjunction, negation, truth and falsum (see (Wansing, 1994) for details). In consequent position, $\bullet A$ may be translated as $\Box A$. We will focus most, however, on the occurrences of $\bullet A$ in a proof system in *antecedent* position (and whenever we talk of the bullet operator from now on, we will assume this interpretation). In antecedent position, it has the intended translation and so the truth condition of a backwards diamond:

$$\mathsf{M}, w \vDash \bullet A \text{ iff } \exists v (Rvw \wedge \mathsf{M}, v \vDash A)$$

**5.2.6 Remark.** There exist several formal translations between all these formalisms. For an overview, we refer to (Lyon et al., 2023). The results include translations between labeled sequents and nested sequents (or tree-hypersequents) (Goré and Ramanayake, 2014; Pimentel, 2018; Lyon, 2021a); translations between labeled and display calculi (Ciabattoni et al., 2021), and between labeled and hypersequent calculi (Lyon et al., 2023), relative to certain logics. Lyon observes that translations from more basic sequents to sequents with a more complex data structure are simpler to define than the other way around. For instance, labeled calculi often only allow for various translations to 'less complex' sequent structures by restricting to labeled sequents in that are in tree form. Although we will not discuss such translations in detail here, we will comment on some of their philosophical implications in Chapter 6 and 7.

### 5.2.1 Restrictions of the hybrid language

The terminology introduced at the beginning of Section 5.2 allows us to consider better the effect of L being extended by an *individual* operator, instead of multiple operators at the same time. In particular, in Chapter 6, this will be relevant for the

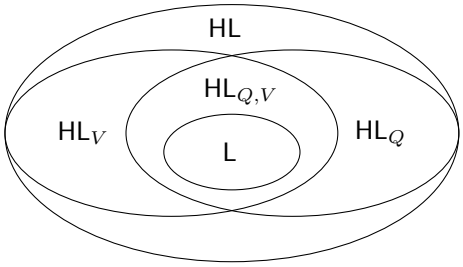



Figure 5.1: The restricted versions of the language HL.

language HL. Considering individual operators is unproblematic in LL, as labels are already only applied to L-formulas (and relational atoms are nullary). In DL, although the display operators can stack, it is only $\bullet A$ that has no translation into L. We thus do not have to make any effort to consider the individual effects of $\bullet$.

However, HL contains multiple hybrid operators $a, @_a, \forall a$ and $\downarrow a$, and the latter three operators are applied to the full language HL. In order to approach the effect of applying these operators just to L, we will try to limit combinations of hybrid operators only to the necessary ones (e.g., in order to have any effect at all, the operators $\forall a$ and $\downarrow a$ must at least combine with name variables $a$). Three restricted versions of HL (see Figure 5.1) will let us distinguish between different context languages for these operators, measuring their effects in different environments. First, consider HL$_V$ ($V$ for 'variable'), which reduces the set of name variables to a singleton.

$$A ::= p \mid a \mid \bot \mid A \wedge A \mid A \to A \mid \Box A \mid @_a A \mid \forall a A \mid \downarrow a A$$
(where $p \in \mathsf{Prop}$, $a \in \mathsf{Var}$ for $|\mathsf{Var}| = 1$)

Second, consider HL$_Q$, which excludes the operator $\forall a A$ from the language ($Q$ for its 'quantifier'-like property).

$$A ::= p \mid a \mid \bot \mid A \wedge A \mid A \to A \mid \Box A \mid @_a A \mid \downarrow a A \quad \text{(where } p \in \mathsf{Prop}, a \in \mathsf{Var})$$

Finally, consider HL$_{Q,V}$, which combines the previous two changes.

$$A ::= p \mid a \mid \bot \mid A \wedge A \mid A \to A \mid \Box A \mid @_a A \mid \downarrow a A$$
(where $p \in \mathsf{Prop}$, $a \in \mathsf{Var}$ for $|\mathsf{Var}| = 1$)

In order to refer conveniently to such language restrictions, we can use the previously defined notions. For example, in the next chapter we might want to consider the formula type $@_a(\mathsf{HL}_V)$, instead of its variant with the natural context language HL.

This way, we will be able to most accurately compare the modal language L to its extensions, without too rashly attributing certain differences to an extended languages as a whole.

## 5.3 Equivalences between (extended) Kripke models

Modal logic is compatible with many types of model-theoretic semantics, including neighborhood semantics, topological semantics, algebraic semantics, and Kripke semantics, the latter of which we already introduced above. For most of the proof-theoretic languages that we defined, regular Kripke models suffice perfectly well for providing truth conditions for their formulas. As we have seen, however, LL and HL require Kripke models to be extended by a label function $\tau$.

In this section, we provide an overview of the notions of equivalence between Kripke models. That is, we want to know when two pointed models are indistinguishable by modal languages: when do they satisfy exactly the same modal formulas? Additionally, we are interested in what a model extension by $\tau$ means for model equivalences. Hence, we will first zoom in on equivalences between regular Kripke models, and then show how these equivalences can be made sense of for extended Kripke models.

### 5.3.1 Equivalences between regular Kripke models

We introduce isomorphisms, disjoint unions, generated submodels and bisimulations for regular pointed Kripke models $(\mathsf{M}, w)$, after their definition in (Blackburn et al., 2001). Consider isomorphisms, which reduce fewest pointed models to each other.

**5.3.1 Definition** (**Isomorphism (**ISO**)**)**.** Two models $\mathsf{M} = (W, R, V)$ and $\mathsf{M}' = (W', R', V')$ are *isomorphic* if there is a function $f : M \to M'$ such that: $w \in V(p)$ iff $f(w) \in V'(p)$; $wRv$ iff $f(w)R'f(v)$; and $f$ is a bijection. Pairs $(w, f(w))$ indicate isomorphic worlds $(\mathsf{M}, w \cong \mathsf{M}', f(w))$.

If $(\mathsf{M}, w)$ and $(\mathsf{M}', w')$ are isomorphic, the frame structure and valuation of $\mathsf{M}$ and $\mathsf{M}'$ are entirely alike. More differences between the models may come in for more tolerant notions of equivalence, that capture better what aspects of Kripke models the modal language cannot express. Disjoint unions, for instance, equate a model with the union of itself and another model.

**5.3.2 Definition** (**Disjoint union** (DU))**.** Two models are disjoint if their domains contain no common elements. For disjoint models $\mathsf{M}_i = (W_i, R_i, V_i)$ $(i \in I)$, their disjoint union is the structure $\biguplus_i \mathsf{M}_i = (W, R, V)$, where $W$ is the union of the sets $W_i$, $R$ is the union of the relations $R_i$, and for each proposition $p$, $V(p) = \bigcup_{i \in I} V_i(p)$. For each $\mathsf{M}_i$, a world $w \in W_i$ is equivalent to its copy in $\biguplus_i M_i$.

On the other hand, generated submodels establish that a model is equivalent to its upward restriction relative to one state.

**5.3.3 Definition (Generated submodel (GS)).** Let $\mathsf{M} = (W, R, V)$ and $\mathsf{M}' = (W', R', V')$ be two models. Then $\mathsf{M}'$ is a *submodel* of $\mathsf{M}$ if: $W' \subseteq W$; $R'$ is the restriction of $R$ to $W'$ (that is: $R' = R \cap (W' \times W')$); and $V'$ is the restriction of $V$ to $\mathsf{M}'$ (that is: for each $p$, $V'(p) = V(p) \cap W'$). We say that $\mathsf{M}'$ is a *generated submodel* of $\mathsf{M}$ if $\mathsf{M}'$ is a submodel of $\mathsf{M}$ and for all points $w$ the following closure condition holds:

$$\text{if } w \text{ is in } \mathsf{M}' \text{ and } Rwv, \text{ then } v \text{ is in } \mathsf{M}'$$

For $\mathsf{M}'$, each world $w \in W'$ is equivalent to its copy in $\mathsf{M}$.

Bisimulations capture the most general model equivalence, reducing the largest number of pointed Kripke models to each other. A bisimulation requires in the most general way that "related states have identical atomic information and matching transition possibilities" (Blackburn et al., 2001).

**5.3.4 Definition (Bisimulation (BI)).** Let $\mathsf{M} = (W, R, V)$ and $\mathsf{M}' = (W', R', V')$. Then $Z \subseteq W \times W'$ is a bisimulation if: if $wZw'$ then $w$ and $w'$ satisfy the same propositional letters; if $wZw'$ and $Rwv$, then there exists a $v' \in M'$ such that $vZv'$ and $Rw'v'$ *(forth)*; and if $wZw'$ and $R'w'v'$, then there exists a $v \in M$ such that $vZv'$ and $Rwv$ *(back)*. Pairs $(w, w')$ related by $Z$ indicate bisimilar worlds $(\mathsf{M}, w \leftrightarrows \mathsf{M}', w')$.

## 5.3.2 Equivalences between extended Kripke models

There are also various ways of adapting the usual notions of equivalence to extended pointed models $(\mathsf{M}, \tau, w)$. These ways generally vary depending on the way that $\tau$ is constrained in the equivalence. In what follows, $\equiv_{\mathsf{P}}$ will denote one of three strengths with which to adapt a usual notion of equivalence to extended pointed models $(\mathsf{M}, \tau, w)$. Consider the weakest extended equivalence.

**5.3.5 Definition (Free extended (FE-)equivalence).** Two extended pointed models $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are *free extended (FE-)equivalent* if their underlying pointed Kripke models are equivalent:

$$\mathsf{M}, \tau, w \equiv_{\mathsf{FE}} \mathsf{M}', \tau', w' \text{ iff } \mathsf{M}, w \equiv \mathsf{M}', w'$$

'Free' indicates that the extended equivalence poses no requirements on $\tau$. This notion extends a regular equivalence simply by adding assignment functions on top of $\mathsf{M}$ and $\mathsf{M}'$. It does not matter what these functions look like: any pair of functions $\tau, \tau'$ added to $\mathsf{M}$ and $\mathsf{M}'$ will lead to the equivalence between $(\mathsf{M}, \tau)$ and $(\mathsf{M}', \tau')$. This is a rather simple way of extending the regular equivalences, but it will provide a proper base comparison to how formulas dependent on name variables stand apart from modal formulas.

Now consider two ways of placing more restrictions on $\tau$. A weak approach connects the object language formulas satisfied at worlds $\tau(a)$ and $\tau'(a)$, and a

strong one explicitly aligns equivalent worlds with the name variables they are assigned.

**5.3.6 Definition (Constrained extended (CE-)equivalence).** Two extended pointed models $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are *constrained extended (CE-)equivalent* if their underlying pointed Kripke models are equivalent, and if object language formulas are invariant under name variable assignments.

$\mathsf{M}, \tau, w \equiv_{\mathsf{CE}} \mathsf{M}', \tau', w'$ if and only if:

1. $\mathsf{M}, w \equiv \mathsf{M}', w'$
2. For all name variables $x$ and $A \in \mathsf{L}$: $\mathsf{M}, \tau(x) \vDash A$ iff $\mathsf{M}', \tau'(x) \vDash A$

'Constrained' indicates that the extended equivalence poses some requirements on $\tau$. It still allows a name variable $a$ to be assigned to non-equivalent worlds. The strongly constrained extended equivalence will not allow this; it is based on *hybrid bisimulations* (Blackburn et al., 2006) (which is intended to provide an invariance result for the basic modal language extended with the satisfaction operator and nominals).

**5.3.7 Definition (Strongly constrained extended (SCE-)equivalence).** Two extended pointed models $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are *strongly constrained extended (SCE-)equivalent* if all states that are assigned a name variable are related by the equivalence, and if equivalent states are assigned the same name variables.

$\mathsf{M}, \tau, w \equiv_{\mathsf{SCE}} \mathsf{M}', \tau', w'$ if and only if:

1. $\mathsf{M}, w \equiv \mathsf{M}', w'$
2. For all name variables $x$, $\mathsf{M}, \tau(x) \equiv \mathsf{M}', \tau'(x)$
3. For all name variables $x$:
   (a) There is a unique $v'$ such that $\mathsf{M}, \tau(x) \equiv \mathsf{M}', v'$, and
   (b) There is a unique $v$ such that $\mathsf{M}, v \equiv \mathsf{M}', \tau'(x)$[6]

'Strongly constrained' indicates that the extended equivalence poses strong requirements on $\tau$. Replacing $\equiv$ in these three definitions of FE-, CE- and SCE-equivalence by one of the equivalences from Section 5.3.1 leads to twelve notions of equivalence between extended models.

Note that strongly constrained extended equivalences require (by the second criterion) that the range of the functions $\tau, \tau'$ is a subset of the states related by the equivalence. For isomorphisms, this is not a problem, because all states of the isomorphic models are by definition already related by the equivalence. For

[6]Thus, the $\tau$- and $\tau'$-ranges of name variables are restricted to being isomorphic. Note that other equivalences can still also have an SCE-version, by the states in $\mathsf{M}$ and $\mathsf{M}'$ that are not assigned any name variables.

disjoint unions, generated submodels and bisimulations, not all states in the model need to be equivalent to some other state: this just means that for SCE-versions of these equivalences, the range of the functions $\tau, \tau'$ need to be restricted to a subset of all states.

In summary, from the above definitions, we can define a more general model equivalence for models for a proof-theoretic language.

**5.3.8 Definition (Equivalences for PL-models).** Let PL be a proof-theoretic language. Then a Kripke model equivalence $\equiv_{\mathsf{P}}$ for PL consists of:

- A regular equivalence $E \in \{\mathsf{ISO}, \mathsf{DU}, \mathsf{GS}, \mathsf{BI}\}$ (as in Section 5.3.1).
- If PL equals LL or HL, a model equivalence extension $S \in \{\mathsf{FE}, \mathsf{CE}, \mathsf{SCE}\}$ (as in Section 5.3.2).

## 5.4 Candidate properties for semantic pollution

We now have the background and formalities that are necessary in order to enter the debate on semantic pollution. In the final part of this chapter, we turn more explicitly towards semantic pollution. After introducing it intuitively, we will briefly consider several properties of proof systems that have intuitive promise to characterize semantic pollution. We check whether these properties are prevalent among calculi that we think are semantically polluted, such as labeled calculi, and scarce among the standard propositional and first-order systems. The properties turn out to not be exactly what we are looking for, and we focus instead on a particular measure acting on the grammar of a proof system in Chapter 6.

### 5.4.1 An introduction to semantic pollution

The phenomenon of semantic pollution has gained increasing attention during the search for satisfactory proof systems for modal logic. Especially labeled calculi are considered to semantically pollute the modal language, by explicitly internalizing Kripke semantics into the proof system. In particular, the calculus introduces labels $x, y, z, ...$ and a 'forcing relation' ':' to accompany every modal formula occurring in the proof system, as well as relational atoms $xRy$ as new primitive formulas. Consider for instance the rules for $\Box$ in the system G3K (Negri, 2005):

$$\frac{y : A, x : \Box A, xRy, \Gamma \Rightarrow \Delta}{x : \Box A, xRy, \Gamma \Rightarrow \Delta} L\Box \qquad \frac{xRy, \Gamma \Rightarrow \Delta, y : A}{\Gamma \Rightarrow \Delta, x : \Box A} R\Box$$

While some proof theorists claim labeled calculi provide desirable technical properties — such as "analyticity, applicability to proof search [and] the possibility to obtain direct completeness proofs" (Negri, 2011) — the literature also refers

to them more apprehensively with philosophical concerns about semantic pollution. Consider for instance that "a philosophical objection to this kind of system is that it builds-in the (desired) semantics into the given syntax" (Braüner and de Paiva, 2006); "[t]he use of a labeled calculus has been sometimes criticized, as mixing semantic elements into what should be a purely syntactic proof system" (Negri, 2011); "some proof-theorists are not satisfied with the idea of labels in proofs that would be seen as 'semantical pollution' because some ingredients of a labeled formalism resemble model-theoretic objects" (Marin, 2018); and "[s]ome have criticised this as a lack of syntactic purity, i.e. as the presence of "semantic pollution"; others defend it as allowing calculi for otherwise unmanageable logics" (Dyckhoff, 2016).

At this point, two main more elaborate philosophical analyses of semantic pollution can be found in the literature. The first is by Read (2015), who argues that the explicit encoding of Kripke semantics in labeled calculi is in fact a virtue, as opposed to tree-hypersequent or nested calculi, which attempt to 'obscure' the semantics in their structural syntax. A recent paper by De Martin Polo (2024) includes a more overarching insight into the types of labeling used by proof theorists, and outlines the current main philosophical attitudes towards using labels. Both authors argue that labeled calculi are suitable for *inferentialism*, the endeavour to specify the meanings of logical constants in terms of their rules of inference — this has been a main potential philosophical drawback of semantically polluted calculi. Thus, the philosophical debate so far seems tentatively ready to accept semantic pollution. However, in the interest of painting a complete picture, we believe that the philosophical views expressing caution with respect to semantic pollution at this point deserve more attention, which we will address in Chapter 6. Additionally, the notion of semantic pollution broadly as a relation between syntax and semantics, remains underspecified in studies so far.

Hence, we have sufficient reason to take a closer look at what semantic pollution could amount to, more formally. As far as we are aware, the only precisification of the notion of syntactic purity (as the counterpart to semantic pollution) has been put forward by Poggiolesi (2010), and appeals to the difference between *internal* calculi and *external* calculi (recall the way we make this precise in Definition 5.2.5). She proposes that a sequent calculus is syntactically pure if it does not "make use of explicit semantic elements", which are exactly elements that make the proof calculus external, i.e., that prevent translation of a sequent to "a formula equivalent to the sequent". Poggiolesi's account follows Avron (1996)'s requirements for 'good' proof systems. There, he suggests that "[a] sequent calculus should be independent of any particular semantic[s]. One should not be able to guess, just from the form of the structures which are used, the intended semantics of a given proof system". Poggiolesi calls this 'strong syntactic purity', since it is hardly satisfied by any proof system: already the classical propositional sequent calculus violates a reasonable interpretation of this type of syntactic purity. Her account then forms a compromised definition of 'weak syntactic purity'.

While the results of this section are largely negative, we get some valuable insights along the way. The first two properties we discuss are categoricity of proof systems and a semantic translation of proof rules — they can both be considered as interpretations of 'guessing' a semantics, and so as interpretations of Avron's strong syntactic purity. For the third property, we will have a look at the notion of 'structural syntax'.

### 5.4.2 Interpretations of 'guessing' a semantics

We first discuss a strict interpretation of 'guessing' a semantics, and one weaker one that is more inspired by Poggiolesi (2010)'s own interpretation of this idea. Let us first generally introduce the relevant properties. We can imagine that "guessing" the intended semantics of a proof system "from the form of the structures which are used" can occur with different levels of certainty. In case of a low level of certainty, the form of the syntactic structures might just induce some idea or impression of similarity to a semantics — but perhaps a high level of certainty implies that the form of the syntax simply fixes the semantics. The latter idea can be captured by categoricity of proof systems.

Categoricity describes the situation where proof rules determine the truth conditions of their connectives. That is, a proof system is categorical when it is only sound and complete with (uniquely specifies) one collection of admissible valuations. Carnap (1943) describes a well-known counterexample, namely the usual proof system for classical propositional logic, which is perfectly consistent with the non-standard valuation that makes everything true (in particular, for each formula $A$, both $A$ and $\neg A$ will come out true, against the truth conditions of negation). Thus, categoricity in the literature is treated as a desirable property, as we want proof calculi to capture the meaning of their connectives and to exclude unwanted interpretations. Note that on this interpretation, semantic influence also becomes a desirable property.

In order to slightly weaken the property of categoricity, we will also talk about categoricity of a *connective*. A connective is categorical when a proof system is categorical — or when a proof system is not categorical, yet non-standard valuations for the particular connective in question are excluded. Thus, the inference rules for a particular connective do determine the semantics of this connective, even though the proof system as a whole is not categorical. This suggests we could take two variations on a requirement for semantic pollution.

**5.4.1 Definition (Candidate property 1).** A proof system is *strongly* semantically polluted if it is categorical. A proof system is *weakly* semantically polluted if it is not categorical, but it has categorical connectives.

The second property of proof systems that we consider is the existence of a way to 'read off' the semantics from the inference rules. A looser interpretation

of 'guessing' the semantics from a proof system can be seen to give rise to this. Given a certain familiarity with the intended semantics of a proof system, the truth conditions of a connective may be recognizable in its inference rules. Such a connection between the proof system and a semantics, instead, might be reason to call a proof system semantically polluted. When analyzing 'strong' syntactic purity, Poggiolesi (2010) gives an interpretation of this idea. She says:

> (Poggiolesi, 2010) [...] the logical rules of $\mathbf{Gcl}_L$ *reflect* at the syntactic level (or may be read in terms of) the semantic definitions of each constant: the elements of the structure of the sequent calculus (i.e. the sequent arrow and the comma) remind us of the metalinguistic elements of the definitions (i.e. *if .. then* and *and* and *or*); the positions of the formulas in the sequent (i.e. the left or the right sides of the sequent) remind us of the truth values in the equivalencies (i.e. false or true).

This idea is made more precise for classical propositional logic by Hacking (1979)'s Do-It-Yourself semantics. Hacking assumes a formal language and a model theory for that language, and presents a way to go from general proof rules to the truth conditions of the connectives defined. His main example concerns sequent rules (omitting side formulas) for a new connective added to classical propositional logic, which occurs in the principal formula $\varphi$:

$$\frac{\{\Gamma_i \Rightarrow \Delta_i\}_{i\in I}}{\Rightarrow \varphi} \qquad \frac{\{\Gamma_j \Rightarrow \Delta_j\}_{j\in J}}{\varphi \Rightarrow}$$

Given such rules, we can tell when $\varphi$ is true or false in a model of the original language: the rule on the left tells us that $\varphi$ is true in a model of the original language iff there are premises $\{\Gamma_i \Rightarrow \Delta_i\}_{i\in I}$ from which to derive $\Rightarrow \varphi$, and if for each premise $\Gamma_i \Rightarrow \Delta_i$ it holds that either some $\gamma \in \Gamma_i$ is false, or some $\delta \in \Delta_i$ is true. The other rule similarly tells us when $\varphi$ is false in a model of the previous language. This inspires the following more general notion of semantic pollution.

**5.4.2 Definition** (**Candidate property 2**)**.** A proof system based in PL is semantically polluted if, given a model theory for PL, there is a *translation*[7] that takes the inference rules for each (existing or new) logical connective to the truth condition of this connective.

While these interpretations provide reasonable conceptions that may be further analyzed future research, we find enough reason to abandon them here. First of all, they are not properties that are reliably absent in the usual proof calculi for propositional or first-order logic, and present in calculi that we think intuitively

[7]There are different ways of making this translation precise: clearly, some preservation of structure is required. However, we will not elaborate on the properties of such a translation, as Hacking's example will already provide reason to reject the measure.

are semantically polluted (especially the labeled calculus). For categoricity, the results in the literature show that the standard calculi for propositional, modal and first-order logic are weakly categorical (suggesting some level of semantic pollution), except somehow intuitionistic propositional systems, which are strongly categorical (suggesting a high level of semantic pollution) (Tong and Westerståhl, 2022). This difference in susceptibility to categoricity does not relate clearly to semantic pollution: the measure of semantic pollution we are looking for would not distinguish the standard classical propositional calculus from its intuitionistic variant. As for the semantic reading of proof rules, this property seems too easily satisfied by proof systems, and simultaneously not developed enough for wide applications. Namely, Hacking relies on strong assumptions for the existence of such a translation method. He presupposes that cut-free derivations are available, and that rules are 'local' in the sense that "they concern only the components from which the principal formula is built up, and place no restrictions on the side formulas" (Hacking, 1979). Hacking's resort to the $\omega$-rule for first-order logic then turns out to enforce classicality (Sundholm, 1981). This means that the "DIY semantics" in practice comes down to quite a restricted notion, and applicable almost only to certain classical propositional systems. And of course, as also noted in the variant of strong syntactic purity of Poggiolesi (2010), the applicability to proof systems for classical propositional logic would tell us exactly that they are examples of semantically polluted calculi. This should already be reason enough to discard the property as a serious candidate for corresponding to semantic pollution as we think of it intuitively.

Future analyses may provide more insight into what methods are necessary to transform a non-categorical proof system, or one that does not allow truth condition translation of its rules, into systems that do allow this — and, whether such methods correspond in any way to proof systems having a more 'semantic nature'. The labeled calculus could serve well as a case study. So far, the literature shows that there are several ways of acting on the proof system in order to ensure categoricity. Among them are ensuring that a proof system has multiple conclusions (Rumfitt, 1997), others are adding a primitive notion of rejection (Smiley, 1996), or working with $n$-sided sequents (Hjortland, 2014). But categoricity can also be ensured by restricting the interpretation space on the semantic side, by imposing certain semantic principles (Bonnay and Westerståhl, 2016). On the syntactic side, increasing expressiveness and changing the structure of sequents is definitely part of what is necessary for categoricity; however, it is unclear whether any of these syntactic extensions can be made sense of as 'semantic' ones. That is, the strengthening of proof systems to multiple-conclusion or bilateralist calculi does not directly involve any referral to semantic notions (such as possible worlds or the Kripke accessibility relation). For multiple-conclusion calculi, the reason that they become categorical seems almost accidental: it is the difference between the empty conclusion being vacuously true for single-conclusion calculi, while it becomes false for multiple conclusions (as truth demands that *some* conclusion

must be true, see also (Rumfitt, 1997)). This seems a technical circumstance that does not seem to hinge on the introduction of multiple conclusions themselves, but more on the choice of their meta-theoretical interpretation. We will thus end the discussion on 'guessing' an intended semantics here, and move on to a different property.

### 5.4.3 Structural syntax

Finally, we discuss the common use of the term 'structural syntax' in relation to semantic pollution. There is a general distinction between structural syntax (and structural rules) and logical syntax (and operational rules) in proof systems. The basic definition of structural syntax in proof systems is the following, found in many places in the literature, and specified to sequents: "A sequent (or rule) is *structural* if it is a schematic statement which does not require mention of any particular connectives" (Humberstone, 2011). Now, the following quote may be seen to relate structure to semantic pollution:

> "In the proliferation of calculi beyond Gentzen systems, there have been two main lines of development, one that enriches the structure of sequents (display calculi, hypersequents, nested sequents, tree-hypersequents, deep inference), another that maintains their simple structure but adds labels and relations in the form of variables and atomic formulas." (Negri, 2016)

Hence, Negri does not consider labels to enrich the structure of sequents: rather, perhaps, they can be seen to enrich the 'operational language' of the proof system.[8] Thus, we might wonder whether the labeled language, as an intuitively semantic language, is 'structural' in a different way than other proof-theoretic languages. We will propose that the definition of structural syntax should be found primarily in their ability to "capture different ways of "bunching data"" (Goré, 1998). This outlook will lead to the modest conclusion that, although in quite a passive way, labels and relational atoms still belong to the structural syntax.

We base our interpretation of structural syntax on the idea that it is more abstract than logical syntax. This is in line with ideas such as that "[o]ur use of the term structural rules [...] is based on the fact that such rules manipulate the underlying data structure of sequents as opposed to introducing more complex logical formulae" (Lyon, 2021c). That is, structural syntax can be thought of as creating different types of distinctions between (collections of) logical formulas, by 'bunching them together' in different ways. While logical connectives also collect formulas together, being structural in the current sense is a one-way street: structure can shape logic, but logic cannot shape structure (merely in the sense that a

[8]Note that we are not claiming that this is a widespread opinion; there are also authors that do classify the labeled language as structural language.

logical operator cannot collect together various structural formulas). Thus, syntax should find itself on a level higher than that of the object language, also related to Došen (1989)'s 'levels' of formulas, inspiring the following simple definitions.

**5.4.3 Definition.** A (possibly nullary) connective $A$ in a proof system is *generally structural* if $A$ cannot be bound by logical connectives. A generally structural connective is *actively structural* if its arity is $\geq 2$. We call syntax *passively structural* if it is generally structural but not actively structural.

Additionally, consider the following notion, that can co-occur with both active or passive structure.

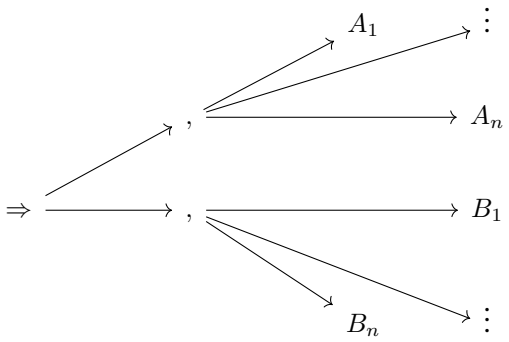
**5.4.4 Definition** ($n$**-level structure**)**.** Let object-language formulas have structural level $0$. A generally structural connective $A$ has *$n$-level structure* (and is of level $n$) if it cannot be bound by formulas of level $n-1$.

We can make simple tree graphs out of any possible complex sequent, by following the arity of the formulas. In the tree graphs that follow, this means that when starting at the root, actively structural connectives change not only the depth of the tree, but also the width (number of branches). The idea behind active structure is that it changes something interesting in how logical formulas are collected together: not only is it of a higher level of abstraction than logical formulas, but it also changes the way that they, or other structure binding logical formulas, are held together. If structure is passive, it acts just as a placeholder of a more abstract nature than what it binds, but that does not 'collect anything in a different way' than the formula that it binds already does. It just adds a 'passive' layer of structure.[9] We can now consider two examples.

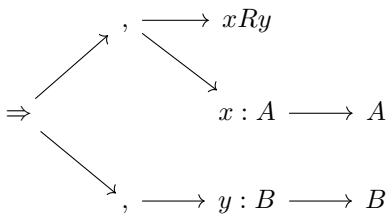
The sequent arrow '$\Rightarrow$' is generally structural, because it cannot be bound by logical connectives. The arrow binds two collections $\Gamma, \Delta$, and so is also actively structural; the Gentzen comma is easily also seen to be actively structural.[10] Furthermore, the comma has $1$-level structure, as it only cannot be bound by logical connectives. The sequent arrow has $2$-level structure, as it cannot be bound by logical connectives, nor by the comma. The following tree visualizes the situation.

[9] Passive here does not mean that the structure is not useful proof-theoretically: it merely means it is passive in the aspect of collecting formulas together.

[10] The comma is $n$-ary in Gentzen systems, but usually displayed explicitly $n-1$ times. For clarity, here we take it as occurring only once and binding $n$ formulas.




Now we are interested in the labeled language. Consider the tree of a labeled sequent $xRy, x : A \Rightarrow y : B$:




Clearly, the labels binding modal formulas are generally structural, but their arity is just unary. Similarly, relational atoms are generally structural, and are nullary. That is, both formula types introduced by the labeled calculus are passively structural. The labeled structure thus essentially adds a structural level to the comma, ensuring that the comma now gets structural level 2, and the sequent arrow structural level 3. A layer in the tree is added for every logical formula, but the labels do not organize logical formulas in a different way. Although the labeled syntax itself codes a graph with its own syntax, this is not reflected in the sequent structure, which we determine by the arity of connectives. In this sense, the structural tree is not in a meaningful way connected to the $R$-graph coded by the labeled syntax.

All in all, we do not take absence or passivity of structuralness as a reliable property indicating semantic pollution. Structural connectives are often intended to mirror logical connectives (see e.g. (Goré, 1998)) on a more abstract level than the logical language. As long as logical connectives can be unary, or nullary, this process does not indicate at all a semantic nature (see for instance the operators $*$ or **I** (interpreted in terms of negation, and truth or falsum) of the display language, which are also passively structural). Additionally, the structural syntactic level can simply also be used to mirror model-theoretic elements: for such 'semantic' structural connectives, perhaps the chance is higher that they are passively structural, because they do not directly relate to or act on the object language. However, passive structuralness by no means guarantees semantic pollution.

## 5.5 Conclusion

This chapter has provided the main preliminaries for our characterization of semantic pollution that will follow next. We are now properly embedded in the existing proof-theoretic languages for modal logic, and in the way that Kripke semantics interprets these languages. We also possess some new terminology in order to focus on the effects of individual operators within a proof-theoretic language. Additionally, we have had a first taste of what semantic pollution is intuitively, and of why several properties for defining semantic pollution more precisely lack relevance. In particular, categoricity of rules, a semantic reading of rules and structuralness of syntax are found not to provide an adequate understanding of semantic pollution. Let us quickly move on to see a more positive characterization of the phenomenon.

# 6

# An ideal of proof systems

## *Characterizing semantic pollution*

This chapter concerns itself with the phenomenon of semantic pollution of proof systems. As we have seen, an intuitive understanding of semantic pollution consists of the idea that a model-theoretic semantics is somehow 'imported' into the proof-theoretic syntax. What exactly semantic pollution amounts to in more concrete terms, however, is largely unclear. Taking Poggiolesi (2010)'s proposal as a starting point, we believe that the wide range of proof systems for modal logic could benefit from a more nuanced account of semantic pollution, and one that is motivated by a more semantic perspective. In attempting to provide one, we will define concrete measures (resulting in four levels) of semantic pollution of proof systems relative to the modal language, and provide a first comprehensive overview of the behaviour of several proof systems for (extensions of) modal logic under these measures. Since the debate on semantic pollution focuses almost entirely on logics with Kripke semantics, we restrict our analysis to this semantics and to proof systems for (extensions of) modal logic. Our framework, however, is intended to be broadly applicable to other logics and semantics, on which we briefly comment in the conclusion as well as Chapter 7. Finally, we embed the general results in a philosophical discussion on semantic pollution. Our categorization is in line with general intuitions about semantically polluted proof systems, providing a formal underpinning for them.

Before we dive into defining measures of semantic pollution, Section 6.1 will provide a few formal remarks and preliminaries to our approach, paving the way to the next sections. A base requirement for semantic pollution will be defined in Section 6.2, which will also analyze several proof systems in terms of this require-

ment. Section 6.3 will then define four levels of semantic pollution by building on the base requirement, and discuss the results. Finally, Section 6.4 will provide an analysis of the philosophical debate surrounding semantic pollution. This chapter corresponds to the main part of the submitted work (Martinot, 2024b).

## 6.1 Preliminary remarks

Our approach has several aspects that are useful to highlight from the beginning.

### 6.1.1 Some formal preliminaries

First of all, semantic pollution will be a property of a proof system, but we will sometimes abuse terminology and say that a formula type or a proof-theoretic language itself is already semantically polluted — but it should be kept in mind that these statements in the end only serve as shorthand for saying that a *proof system* is semantically polluted (one based in the particular proof-theoretic language and formula type).

Furthermore, it will not be possible for $\mathsf{L}$-formulas to be semantically polluted: the same will hold for proof-theoretic formulas *translatable* to $\mathsf{L}$ (as defined in the previous chapter). Hence, if the translatability of a proof-theoretic formula depends on the background logic, then whether proof-theoretic formulas are semantically polluted does, too. This for instance concerns common extensions of the Gentzen calculus for modal logic, like the nested calculus or tree-hypersequent calculus, and the hypersequent calculus (and even the usual Gentzen arrow and comma). Relative to a background logic, these structural sequents commonly have intended translations in terms of, for instance, disjunction and the box operator. Our approach will render them automatically syntactically pure. However, such calculi have been speculated to possess (some version of) semantic pollution by Read (2015); De Martin Polo (2024); hence, we will discuss them more in Section 6.4.

For defining the results of semantic pollution, the restricted versions of $\mathsf{HL}$ as defined in Chapter 5 will come in handy. Recall that this is because we are mainly interested in the effects of extending $\mathsf{L}$ by *individual* operators (that are, as far as possible, only applied to formulas in $\mathsf{L}$), so that we can more easily compare their individual behaviour to that of $\mathsf{L}$-formulas, and to that of other proof-theoretic operators. The restricted versions of $\mathsf{HL}$ (see Figure 5.1) will let us distinguish between different semantic pollution results, that will indicate from which language they are obtained.

More specifically, throughout the chapter, we will propose a base requirement for semantic pollution, and four more elaborated levels of semantic pollution. Satisfaction of the base requirement and of a level of pollution will then be introduced to a proof-theoretic language, and in turn to a proof system, by a *formula type*. By

focusing on formula types, the truth conditions of proof-theoretic formulas will be analyzed at the level of their main operator. This way, it is really the operators introduced by the proof system that get full responsibility for their semantic effects.

The results of the base requirement come with some notation. In general, satisfaction of the base requirement (written as BR) will be indicated by the abbreviation 'po' (for pollution, saying a formula type satisfies BR), while syntactic purity results will be indicated by the abbreviation 'pu' (for purity, saying a formula type does not satisfy BR). For the language HL, we will furthermore distinguish between results from different context languages, providing several levels of satisfying BR. Now suppose we are given the definition for BR. Then its satisfaction is determined for a formula type $C(\mathsf{L})$ with respect to the *modal* context language L.

- The result po means $C(\mathsf{L})$ satisfies BR (po for (semantic) pollution).[1]
- The result pu means that $C(\mathsf{L})$ does not satisfy BR (pu for (syntactic) purity).

For HL, $C(\mathsf{L})$ sometimes captures 'vacuous purity results'. Namely, its operators cannot always show their power when restricted to L. On the other hand, $C(\mathsf{HL})$ will sometimes fail to show the *individual* effect of $C$, as $C$ can combine with any other operator of HL. Hence, we will introduce some weaker types of pollution that arise for context languages $\mathsf{L}'$ such that $\mathsf{L} \subset \mathsf{L}' \subseteq \mathsf{HL}$. po is then the strongest type of pollution for HL-operators. A result $\mathsf{po}_{\mathsf{L}'}$ will indicate that $\mathsf{L}'$ is the smallest language (among the ones we have selected) to satisfy BR. This means that the operator restricted to a smaller language will not satisfy BR, and will instead satisfy a purity result: several of such results will be proven in the Appendix. We use the following notation for weaker types of pollution.

- The result $\mathsf{po}_{Q,V}$ means that $C(\mathsf{HL}_{Q,V})$ satisfies BR.
- The result $\mathsf{po}_{V}$ means that $C(\mathsf{HL}_{V})$ satisfies BR.
- The result $\mathsf{po}_{Q}$ means that $C(\mathsf{HL}_{Q})$ satisfies BR.
- The result $\mathsf{po}_{\mathsf{HL}}$ means that $C(\mathsf{HL})$ satisfies BR.

It will follow from the definition of BR that po implies $\mathsf{po}_{Q,V}$. Furthermore, $\mathsf{po}_{Q,V}$ implies $\mathsf{po}_{V}$ and $\mathsf{po}_{Q}$, and the latter two both imply $\mathsf{po}_{\mathsf{HL}}$, which is the weakest type of pollution.

Finally, in this chapter, the general notation PL for proof-theoretic language will have just three instantiations, as defined in the previous chapter:

- The labeled language (LL)
- The display language (DL)
- The hybrid language (HL)

[1] We will assume that $\bullet(\mathsf{L}) = \bullet(\mathsf{DL})$, as their results for the measures for semantic pollution will be the same.

### 6.1.2 Conceptual preliminaries

Our method has a few characteristic conceptual properties, which we here emphasize for clarity. First, our approach is heavily *framework-dependent*. We see semantic pollution as dependent on an object language, a proof-theoretic language, and a particular semantics. We thus do not accommodate a notion of semantic pollution that is irrespective of this background context, and that only relies on proof-theoretic tools. We consider this to be natural: there is simply no absolute standard for syntax to be 'syntactic' (any expression can in principle included in a syntax) — hence, we need a baseline connection between an object language and a semantics in order to distinguish 'semantic' and 'syntactic' syntax in the proof-theoretic language.

Second, we define *'operator-level'* semantic pollution (by using formula types). One might instead focus on properties of individual concrete formulas (members of a formula type), or even of languages as an entirety. However, the interesting level of detail seems to concern operators that a proof-theoretic language introduces. This emphasizes that the object language itself is syntactically pure, and that the proof system introduces pollution by *adding* to this language.

Third, we see semantic pollution as something *static*: if a formula type satisfies it, then the proof system as a whole (and any formal proof using instances of the formula type) can be considered semantically polluted. This perspective ignores the behaviour of proof-theoretic syntax inside inference rules (i.e., the way formulas are *used* in proofs). For instance, if the use of a semantic formula in a proof is easily eliminable, this might take away from the level of semantic pollution. However, this encourages a minimal view of semantic pollution, as there often exist many translations between proof systems, where 'semantic' properties may be lost. We prefer to consider proof systems individually, and to evaluate their design. And perhaps especially when the uses of a 'semantic' formula type in a proof system are easily eliminable, the proof system should be considered semantically polluted: why should one introduce semantic notions into a language, if it is not even necessary?[2]

**6.1.1 Remark.** There is a general distinction between 'bottom-up' and 'top-down' approaches to formalizing philosophical notions. Generally, bottom-up approaches focus on the use of this notion by experts in practice, and aim to provide a formalization that matches this practice closely. Top-down approaches, on the other hand, develop a framework based on theoretical (possibly idealized) principles intuitively underlying a notion. Practical examples may then instead be measured by the standard of this framework. Both approaches are valuable for different reasons. This paper starts 'bottom-up' by taking the intuitions on semantic pollution mentioned in Section 5.4.1 as a given; we aim to provide a formal framework capturing the idea that labeled calculi possess most semantic pollution, and that the

[2] See for more comments on this topic Section 6.4.2.

usual (e.g. propositional) sequent calculus possesses none. The inspiration for and specification of our measures of semantic pollution in Section 6.2 and 6.3, however, also come with top-down influences on what we think makes up a 'semantic nature'.

## 6.2 The base requirement: violating invariance results under model equivalences

This section will define the base requirement of semantically polluted formula types. By itself, this property provides enough information to conclude that a formula type is semantically polluted, and it specifies a level of pollution with respect to it. However, the next section will introduce two properties on top of the one defined here, that more clearly divide the different formula types among four levels of semantic pollution. The idea of the base requirement comes in when we consider the connection that the (classical) basic modal language $\mathsf{L}$ displays towards Kripke semantics as a syntactically pure baseline. That is, we can view the way that the basic modal language distinguishes Kripke models, by describing Kripke models with a certain level of detail, as a syntactically pure standard. The basic modal language partitions the space of pointed Kripke models based on what it can express about these models, by equating pairs $(\mathsf{M}, w)$ that it considers to be the same (i.e., that cannot be distinguished by formulas of $\mathsf{L}$).

Formulas of a proof-theoretic language extending $\mathsf{L}$ may then be found to make *more* distinctions between pointed Kripke models than $\mathsf{L}$, and create a more fine-grained partition. This tells us that the formula can express differences between two (modally equivalent) worlds in two models that no modal formula can. It can describe a Kripke model in a way that is unavailable to the modal language — and in this sense, the formula has a stronger connection to the semantics. We can check this by considering whether a proof-theoretic formula violates invariance results under Kripke model equivalences of $\mathsf{L}$. As mentioned before, note that this is a semantic perspective and elaboration of the suggestion proposed in (Poggiolesi, 2010) that semantic pollution comes down to untranslatability to an object language. Namely, violating invariance results under model equivalences for $\mathsf{L}$, is a way of establishing untranslatability to $\mathsf{L}$. We will now formally define the property of violating invariance results under Kripke model equivalences.

### 6.2.1 Levels of satisfying the base requirement

There are three aspects affecting the level of satisfying the base requirement for formula types from $\mathsf{HL}$, two for the formula types from $\mathsf{LL}$ and one for the formula type of $\mathsf{DL}$.

**Model equivalence (DL, LL, HL).** This concerns baseline notions of equivalence for Kripke models. Generally, it is most difficult for a formula type to violate invariance results under model equivalences that reduce many models to each other, and this will indicate a higher level of pollution. We will use the symbol $\equiv$ as covering two notions of equivalence as defined in Chapter 5. That is, $\mathsf{M}, w \equiv \mathsf{M}', w'$ means either:

1. $(\mathsf{M}, w)$ is *isomorphic* to $(\mathsf{M}', w')$, i.e. $\mathsf{M}, w \cong \mathsf{M}', w'$
2. $(\mathsf{M}, w)$ is *bisimilar* to $(\mathsf{M}', w')$, i.e. $\mathsf{M}, w \leftrightarroweq \mathsf{M}', w'$

Note that although we can make more distinctions by picking more notions of equivalence in between isomorphisms and bisimulations (such as generated submodels and disjoint unions), we believe too many equivalences will only obscure the results, and these two extremes already form a suitable representation of available equivalences and lead to variable results.

In particular, when considering isomorphisms, violating invariance results can only happen if a formula does something independent from the local frame and valuation structure that surrounds $w$ and $w'$. For instance, unconstrained assignment functions $\tau$ can scan different parts of the same model, leading to different truth values of formulas including labels or nominals.

Although we do not include disjoint unions and generated submodels, note that potential reasons for violating invariance under one of these notions increase, as they reduce more models to each other. In disjoint unions, formulas that say something about the global model situation can additionally cause violation of invariance, since a disjoint union adds an entire model to an initial one, and a formula may be able to express this. In generated submodels, a formula may furthermore detect that local worlds preceding equivalent worlds disappear.

Both of these aspects are also captured by bisimulations. A formula may additionally violate invariance under bisimulations because it detects differences in the number of local successors of equivalent worlds, which can vary under bisimulation. Thus, bisimulations capture more types of semantic pollution than isomorphisms, which are more strict in what counts as pollution.

**Model equivalence extension (LL, HL).** This concerns the extension of the two model equivalences to models with an assignment function. We will take the three 'equivalence strengths' for extended models $(\mathsf{M}, \tau)$ as defined in Chapter 5, corresponding to the following notation:

1. Free extended (FE-)equivalence $\equiv_{\mathsf{FE}}$.
2. Constrained extended (CE-)equivalence $\equiv_{\mathsf{CE}}$.
3. Strongly constrained extended (SCE-)equivalence $\equiv_{\mathsf{SCE}}$.

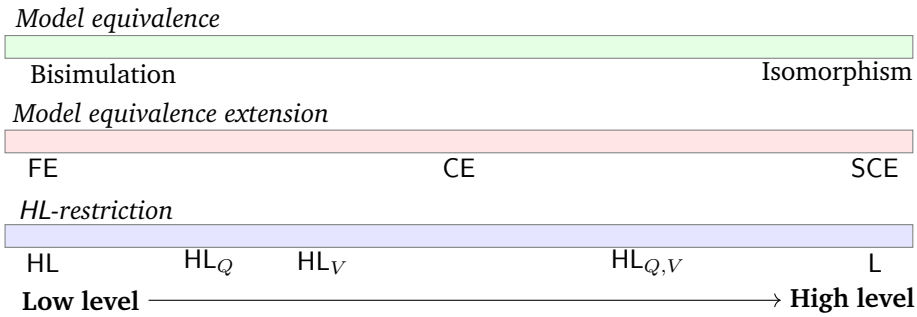



Figure 6.1: Three aspects affecting the level of satisfying the base requirement. 'Model equivalence' and 'model equivalence extension' apply to LL, DL and HL, while 'HL-restriction' applies only to HL.

For free-extended equivalences, note that violations of invariance results by formulas including name variables may be rather unsurprising, since $\tau$ and $\tau'$ can map name variables to very different states in the equivalent models. Lack of surprise does not indicate lack of value, however: such results show exactly that $\tau$ scans a Kripke model in a way that is foreign to the modal language, a phenomenon that we aim for semantic pollution to capture.

The two constrained equivalences will indicate a higher level of semantic pollution, as a formula violating invariance results under these equivalences distinguishes even more Kripke models than formulas only violating invariance results under FE-equivalences. That is, it is more difficult for a formula type to violate invariance results under a stronger equivalence for extended models.

We will use the symbol $\equiv_{\mathsf{P}}$ as in Definition 5.3.8, only now restricted to only *two* regular equivalences as defined above, combined with the three possible equivalence strengths.

**Size of the context language (HL).** It is most difficult for an HL-formula type to violate invariance results relative to a small context language (see Figure 5.1), and so this will indicate a higher level of pollution. Note that this does not apply to LL and DL.

Figure 6.1 visualizes the effect of these three properties on satisfying the base requirement. Now we will see the model equivalences and model equivalence extensions defined in detail.

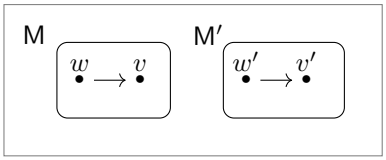



Figure 6.2: Example of the base requirement.

### 6.2.2 Base requirement for semantic pollution

The work of the previous sections now comes together in the base criterion that any semantically polluted formula type will need to satisfy.

**6.2.1 Definition** (**Base requirement (**BR**)**)**.** Let $C$ be an operator or metavariable of PL, and let CL be the *context language* of $C$. Let $\equiv_{\mathsf{P}}$ consist of an extended Kripke model equivalence with the following ingredients:

- A regular equivalence $E \in \{\mathsf{ISO}, \mathsf{BI}\}$ (as in Section 5.3.1)
- If PL equals LL or HL, a model equivalence extension $S \in \{\mathsf{FE}, \mathsf{CE}, \mathsf{SCE}\}$ (as in Section 5.3.2)

Then $C$ satisfies $\mathsf{BR}_{\mathsf{CL},E,S}$, i.e., the *base requirement of semantic pollution relative to context language CL, equivalence $E$ and strength $S$* if it violates invariance under $\equiv_{\mathsf{P}}$:

There are equivalent pointed models $\mathsf{PM}, w \equiv_{\mathsf{P}} \mathsf{PM}', w'$, and an $A \in C(\mathsf{CL})$ such that

1. $\mathsf{PM}, w \vDash A$
2. $\mathsf{PM}', w' \nvDash A$

This requirement gives us a level of pollution with respect to the aspects shown in Figure 6.1. We will discuss the results of BR with respect to the formula types of DL, LL and HL in the next section, using the variants for pollution (po) and purity (pu) as described in Section 6.1. In order to tear apart the formula types even more clearly, we will define more overarching levels of semantic pollution in Section 6.3. First, a small example of BR for an intuitive sense of it.

**6.2.2 Example** (**Example of the base requirement**)**.** Take the operator $@_a$ of HL, and suppose we want to show that it satisfies $\mathsf{BR}_{\mathsf{L},\cong,\mathsf{FE}}$. Figure 6.2 shows two models $(\mathsf{M}, \tau)$ and $(\mathsf{M}', \tau')$, where we let $V(p) = w$, $V'(p) = w'$, $\tau(a) = w$, $\tau'(a) = v'$. Then $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are FE-isomorphic, yet $\mathsf{M}, \tau, w \vDash @_a p$ and $\mathsf{M}', \tau', w' \nvDash @_a p$. Hence, $@_a(\mathsf{L})$ satisfies $\mathsf{BR}_{\mathsf{L},\cong,\mathsf{FE}}$.

Finally, we note that the base requirement is theoretically applicable to any other combination of object language and model-theoretic semantics. However,

| **Equivalence** | $\bullet A$ |
|---|---|
| Isomorphism | pu |
| Bisimulation | po |

Table 6.1: Base requirement results of the bullet operator.

the details of making the requirement precise depend too much on the particular context for a useful generalization at this point.

### 6.2.3 Results

We treat the results per language DL, LL and HL. Most examples we provide are well-known and can already be found elsewhere, for instance in (Blackburn et al., 2001), but we discuss them here for the first time within the context of semantic pollution.

**Results of** DL**.** From DL, we see that only $\bullet A$ as interpreted in the antecedent position of a sequent (see Section 5.2) satisfies the base requirement. By their translatability to L, **I**, $A \circ A$ and $*A$ stay syntactically pure.

$\bullet A$ gives us simple results (see Table 6.1). Results for extended model equivalences do not apply to it, so we only consider the usual variants of Kripke model equivalences. There, we see that it only violates invariance under bisimulations, simply because it functions like a backwards diamond. This also indicates only a medium level of semantic pollution, as isomorphisms preserve its truth value, and they are harder to violate invariance under (note that a similar pattern holds if we included generated submodels and disjoint unions — $\bullet A$ is polluted with respect to the former, but pure with respect to the latter).

**Results of** LL**.** Consider now labeled formulas and relational atoms, that also give us rather clear-cut results (see Table 6.2). Interestingly, their semantic pollution is indifferent to the type of Kripke model equivalence chosen (bisimulation or isomorphism), for all model equivalence extensions. This indicates that $\tau$ rises above the differences in model properties that the modal language cannot see. Consider the following example for FE-isomorphisms (and so also bisimulations).

**6.2.3 Example** ($x : p$, $xRy$)**.** Consider the models in Figure 6.2, and let $V(p) = \{w\}$, $V'(p) = \{w'\}$. Let $\tau(x) = w$, $\tau'(x) = v'$, $\tau(y) = v$ and $\tau'(y) = v'$. Then $(\mathsf{M}, \tau, w)$ is FE-isomorphic (and observe, not CE- or SCE-isomorphic) to $(\mathsf{M}, \tau', w')$. However, $\mathsf{M}, \tau, w \vDash x : p$, $\mathsf{M}', \tau', w' \nvDash x : p$, and $\mathsf{M}, \tau, w \vDash xRy$, $\mathsf{M}', \tau', w' \nvDash xRy$.

Stronger model equivalence extensions reduce pollution, and also tear apart $x : A$ and $xRy$, as relational atoms are more semantically polluted than labeled

| **Model equivalence extension** | $x : A$ | $xRy$ |
|---|---|---|
| Free extended | po | po |
| Constrained extended | pu | po |
| Strongly constrained extended | pu | pu |

Table 6.2: Base requirement results of LL-operators (results are the same for isomorphisms and bisimulations of the same model equivalence extension).

| **Model equivalence extension** | $a$ |
|---|---|
| Free extended | po |
| Constrained extended | po |
| Strongly constrained extended | pu |

Table 6.3: Base requirement results of $a$ (results are the same for isomorphisms and bisimulations of the same model equivalence extension).

formulas. Of course, this is by design of the model equivalence extensions: CE-equivalences purify $x : A$ by definition, while they do not yet tie down $xRy$ (just edit the previous example by giving all worlds valuation $\{p\}$ — just because each world satisfies the same modal formulas, does not mean they have the same $R$-context).

Only SCE-equivalences tame $xRy$. To see this, suppose that $\mathsf{M}, \tau, w \vDash xRy$, and suppose there is a strong equivalence $\mathsf{M}, \tau, w \equiv \mathsf{M}', \tau', w'$. By assumption, $\tau(x)R\tau(y)$, and we know that the entire range of $\tau$ is included in the equivalence. Hence, there are worlds $s'$ and $t'$ in $\mathsf{M}'$ such that $\tau(x) \equiv s'$ and $\tau(y) \equiv t'$. We also know that equivalent worlds must satisfy the same name variables, so $s' = \tau'(x)$ and $t' = \tau'(y)$. Now to see that $s'Rt'$, suppose that our equivalence is a bisimulation (for isomorphisms, $s'Rt'$ is clear). By the forward condition, there exists $u' \in \mathsf{M}'$ such that $s'Ru'$ and $\tau(y) \equiv u'$. But as equivalent worlds must satisfy the same labels, $u' = \tau'(y)$, and so $u' = t'$, so that $s'Rt'$ (and $\mathsf{M}', \tau', w' \vDash xRy$).

**Results of HL.** The results of HL are a little more intricate. First consider the simple results of $a$ as in Table 6.3, and note that it has the same results as $xRy$. The following example illustrates the cases of pollution.

**6.2.4 Example** ($a$)**.** Take Figure 6.2 and let all worlds have valuation $\{p\}$. Let $\tau(a) = w$ and $\tau'(a) = v'$. Then $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are FE- and CE-equivalent. Yet $\mathsf{M}, \tau, w \vDash a$ and $\mathsf{M}', \tau', w' \nvDash a$.

Clearly, strongly constrained extended equivalences will purify $a$ by definition, because of the strict demand that equivalent worlds are assigned the same name variables.

We then switch to the more graded results of the operators $@_a$, $\forall a$ and $\downarrow a$. Re-

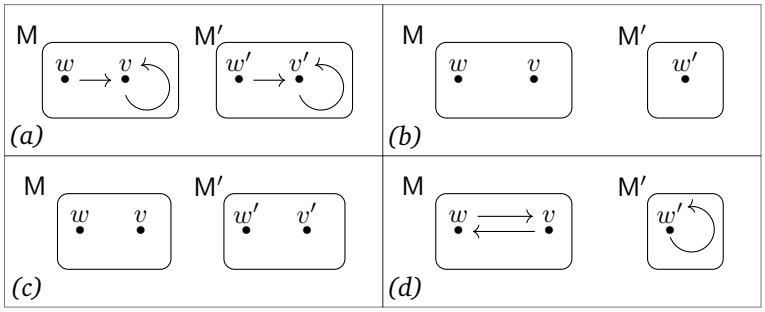



Figure 6.3: The extended models serving as proof for satisfying the base requirement. Assume that all worlds receive valuation $\{p\}$.

| FE-**equivalence** | $@_a A$ | $\forall a A$ | $\downarrow a A$ |
|---|---|---|---|
| Isomorphism | po | po($\mathsf{HL}_Q$) | po($\mathsf{HL}_Q$) |
| Bisimulation | po | po($\mathsf{HL}_{Q,V}$) | po($\mathsf{HL}_{Q,V}$) |

Table 6.4: Base requirement results of HL-operators for free extended equivalences.

call that we consider them with respect to several language fragments. Note from Table 6.5 that $@_a : A$ shows more pollution than $x : A$, as it retains pollution in a context language more powerful than L. The difference is seen in CE-equivalences, where $x : A$ is immediately turned pure, but $@_a A$ finds pollution even in the rather restricted language $\mathsf{HL}_{Q,V}$: consider the example below.

**6.2.5 Example** ($@_a \Diamond a$)**.** Consider the pair of models in Figure 6.3(a). Let $\tau(a) = w$ and $\tau'(a) = v'$. Then $(M, \tau, v)$ is CE-isomorphic to $(M', \tau', v')$. Yet $M, \tau, v \not\vDash @_a \Diamond a$, and $M', \tau', v' \vDash @_a \Diamond a$.

In a strongly constrained extended isomorphism (see Table 6.6), $@_a$ is finally turned pure. As Theorem 6.6.4 shows, in fact, all formulas in HL are invariant under SCE-isomorphisms.

But SCE-bisimulations still show a low level of pollution for $@_a A$ relative to $\mathsf{HL}_V$, by interacting with $\forall a A$. Consider the example below.

**6.2.6 Example** ($@_a(\forall a(a))$)**.** Consider the pair of models in Figure 6.3(b). Let $\tau(a) = w$ and $\tau'(a) = w'$. Then $(\mathsf{M}', \tau', w')$ is SCE-bisimilar $(\mathsf{M}, \tau, w)$. Yet $\mathsf{M}, \tau, w \not\vDash @_a \forall a(a)$, while $\mathsf{M}', \tau', w' \vDash @_a \forall a(a)$.

To see that for bisimulations, this is the highest level of pollution that $@_a A$ can have, Theorem 6.6.5 shows that all hybrid formulas in $\mathsf{HL}_Q$ are invariant under SCE-bisimulations. Thus, any smaller language than $\mathsf{HL}_V$ will purify $@_a A$.

Next, consider the operator $\forall a A$ separately. Just like in the example above,

| CE-**equivalence** | $@_a A$ | $\forall a A$ | $\downarrow a A$ |
|---|---|---|---|
| Isomorphism | $\mathsf{po}(\mathsf{HL}_{Q,V})$ | $\mathsf{po}(\mathsf{HL}_{Q})$ | $\mathsf{po}(\mathsf{HL}_{Q})$ |
| Bisimulation | $\mathsf{po}(\mathsf{HL}_{Q,V})$ | $\mathsf{po}(\mathsf{HL}_{Q,V})$ | $\mathsf{po}(\mathsf{HL}_{Q,V})$ |

Table 6.5: Base requirement results of HL-operators for constrained extended equivalences.

| SCE-**equivalence** | $@_a A$ | $\forall a A$ | $\downarrow a A$ |
|---|---|---|---|
| Isomorphism | pu | pu | pu |
| Bisimulation | $\mathsf{po}(\mathsf{HL}_{V})$ | $\mathsf{po}(\mathsf{HL}_{Q,V})$ | $\mathsf{po}(\mathsf{HL}_{V})$ |

Table 6.6: Base requirement results of HL-operators for strongly constrained extended equivalences.

$\forall a$ is able to make restricted cardinality statements in a small language, and so is easily able to continuously satisfy the base requirement for non-isomorphism equivalences. An easy example works for bisimulations of all three strengths (in the language $\mathsf{HL}_{Q,V}$, accounting for the results in Table 6.4, 6.5, 6.6): simply take Example 6.2.6 for just the formula $\forall a(a)$. This is why the level of SCE-pollution of $@_a$ relative to $\mathsf{HL}_V$ has to be seen as rather low (it is mainly 'caused by' $\forall a A$). Note also that these results for bisimulations indicate the highest level of pollution that $\forall a A$ can achieve, as it is invariant under L (the only smaller language than $\mathsf{HL}_{Q,V}$).

For isomorphisms, there is an example in the restricted language $\mathsf{HL}_Q$ that still works for $\forall a A$ in the FE and CE case (see Table 6.4, 6.5).

**6.2.7 Example** ($\forall a(a \vee b)$)**.** Consider the pair of models in Figure 6.3(c). Let $\tau(a) = w$, $\tau(b) = w$, $\tau'(a) = w'$, $\tau'(b) = v'$. Then $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are both FE- and CE-isomorphic. But $\mathsf{M}, \tau, w \vDash \forall a(a \vee b)$, while $\mathsf{M}', \tau', w' \nvDash \forall a(a \vee b)$.

The fact that $\mathsf{po}(\mathsf{HL}_Q)$ is the highest level of pollution of $\forall a A$ for FE-isomorphisms, is shown by Theorem 6.6.3 (showing that $\forall a A$ is invariant under FE-isomorphisms relative to $\mathsf{HL}_V$). This result extends to CE-isomorphisms, which only pose more requirements on $\tau$. However, as mentioned before, Theorem 6.6.4 shows that SCE-isomorphisms purify $\forall a A$ (see Table 6.6).

Finally, consider $\downarrow a A$. By resorting to reflexivity statements, it can satisfy the base requirement relative to FE- and CE-bisimulations and $\mathsf{HL}_{Q,V}$ (Table 6.4, 6.5), as in the following example.

**6.2.8 Example** ($\downarrow a(\Diamond a)$)**.** Consider the pair of models in Figure 6.3(d). Let $\tau(a) = w$ and $\tau'(a) = w'$. Then $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are FE- and CE-bisimilar. Yet $\mathsf{M}, \tau, w \nvDash \downarrow a(\Diamond a)$, while $\mathsf{M}', \tau', w' \vDash \downarrow a(\Diamond a)$.

Just like for $\forall a A$, note that $\mathsf{po}(\mathsf{HL}_{Q,V})$ is the highest level of pollution for these

categories, as $\downarrow aA$ is pure relative to L (the only smaller language than $\mathsf{HL}_{Q,V}$). For FE- and CE-isomorphisms, the following polluting example of $\downarrow aA$ is taken from $\mathsf{HL}_Q$, a less restricted language, and so less polluting (Table 6.4, 6.5).

**6.2.9 Example** ($\downarrow a(@_b a)$)**.** Consider the pair of models in Figure 6.3(c). Let $\tau(a) = w$, $\tau'(a) = w'$, $\tau(b) = w$, and $\tau'(b) = v'$. Then $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are FE- and CE-isomorphic, yet $\mathsf{M}, \tau, w \vDash \downarrow a(@_b a)$, while $\mathsf{M}', \tau', w' \nvDash \downarrow a(@_b a)$.

To see that this $\mathsf{po}(\mathsf{HL}_Q)$ result is the highest level of pollution for FE-isomorphisms, see Theorem 6.6.2 (showing invariance of $\downarrow aA$ under FE-isomorphisms relative to $\mathsf{HL}_V$). Again, this result extends to CE-isomorphisms. Finally, an example using $\forall aA$ comes back in, and gives a low pollution result for SCE-bisimulations relative to $\mathsf{HL}_V$ (Table 6.6).

**6.2.10 Example** ($\downarrow a(\forall a(a))$)**.** Consider the pair of models in Figure 6.3(b). Let $\tau(a) = w$ and $\tau'(a) = w'$. Then $(\mathsf{M}, \tau, w)$ and $(\mathsf{M}', \tau', w')$ are SCE-bisimilar, yet $\mathsf{M}, \tau, w \nvDash \downarrow a(\forall a(a))$, while $\mathsf{M}', \tau', w' \vDash \downarrow a(\forall a(a))$.

Theorem 6.6.5 then shows that (among others) $\downarrow aA$ is invariant under SCE-bisimulations relative to $\mathsf{HL}_Q$, implying that $\mathsf{po}(\mathsf{HL}_V)$ is indeed the highest pollution level here for $\forall aA$. And once more, the purity result of $\forall aA$ for SCE-isomorphisms relative to HL is shown by Theorem 6.6.4.

### 6.2.4 Summing up

The base requirement for semantic pollution can be satisfied with various levels, depending on model equivalence, model equivalence extension (if applicable) and context language restriction (if applicable). These aspects already create a division in satisfaction of the base requirement for formula types in LL, DL and HL. The results are summarized in Figure 6.4. The lowest level of satisfaction of the base requirement is steadily provided by $\bullet A$. The highest level is displayed by $xRy$ and $a$, which are only 'purified' for strongly constrained extended equivalences, showing the independence of $\tau$ from the basic modal language.

The level of semantic pollution of $@_a A$ is reduced significantly by strong equivalences, but never completely eliminated (unlike $x : A$), due to its interaction with other formulas in HL. The level of BR-satisfaction for $\downarrow aA$ and $\forall aA$ is relatively low, yet (especially for $\forall aA$) remains persistent throughout the different model equivalence extensions. Only for SCE-equivalences does $\downarrow aA$ get similar results to $@_a A$, and $\forall aA$ retains the highest level of pollution there. However, although low pollution levels remain with SCE-equivalences for $@_a A$, $\downarrow aA$ and $\forall aA$, note that these all rely on the workings of $\forall aA$. That is, they rely on the ability of $\forall$

[3]Two formula types have a different level of pollution in the figure, if this difference exists with respect to isomorphisms, bisimulations, or both. Additionally, although $\bullet A$ does not need FE-, CE- and SCE-equivalences, its results can be seen as unchanged with respect to these notions.

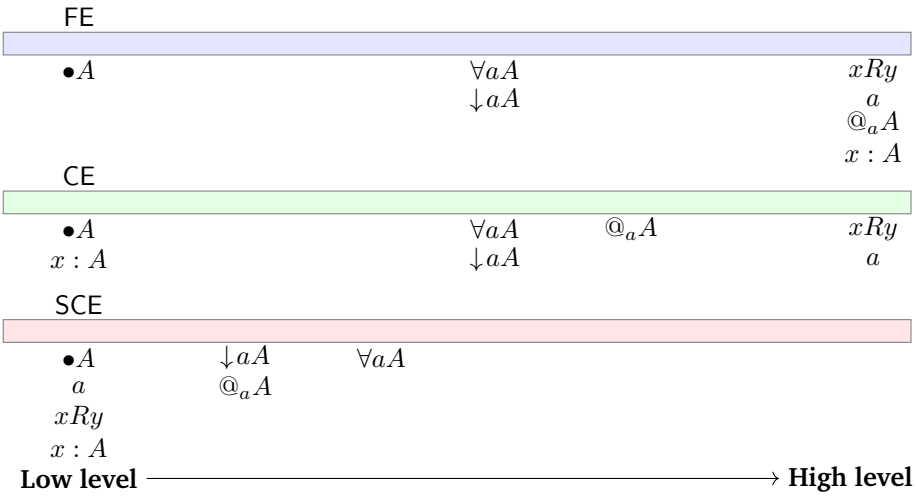



Figure 6.4: Summary of the results for (levels of) satisfaction of BR.[3]

to distinguish between a model cardinality of one and more than one. Clearly, we should emphasize that for $@_aA$ and $\downarrow aA$ this is a side effect of the context language $\mathsf{HL}_V$, and this is a low level of BR-satisfaction. For $\forall aA$ itself, the property is more inherent and reflective of a proper remaining level of semantic pollution (relative to the minimal context language $\mathsf{HL}_{Q,V}$).

Furthermore, we see that the BR-results of $\downarrow aA$ for SCE-bisimulations hold with respect to a larger context language than those of $\forall aA$. This shows that sending $a$ to the current world (as $\downarrow aA$ does) is something less invasive than sending it to *all* worlds (as $\forall aA$ does). Both operators change the assignment function, but $\downarrow aA$ changes it only by using local information, while $\forall aA$ changes it by using global information. Note that both formulas are still syntactically pure for isomorphisms. Thus, there must be some 'actual' difference between two equivalent models for $\forall aA$ and $\downarrow aA$ to pick up on. This is unlike more direct uses of the assignment function $\tau$ as in $xRy$, $x : A$ or $a$, which can more easily vary even for very similar models, but are also more easily be constrained again by stronger equivalences.

So far, we provided a general distinction between levels of BR-satisfaction. In the next section, we will highlight two aspects that should, according to us, count more strongly in determining the exact level of pollution. This provides a more intuitive division into various types of semantic pollution, and emphasizes the differences that the base requirement cannot capture.

## 6.3 Four levels of semantic pollution

The base requirement for semantic pollution is insightful, and distinguishes between some degrees of semantic pollution. However, we believe that the emphasis of two properties (underlying some of the results in the base requirement) should give rise to stronger forms of semantic pollution. They can be interpreted as being 'less modal' and 'more semantic' in two respects, and thus form more of an unnatural invasion into the modal language than lower forms of semantic pollution.

**Globalness.** The first is the property of globalness, or world invariance. It is well-known that "[m]odal satisfaction is intrinsically local: only the points accessible from the current state are relevant to truth or falsity" (Blackburn et al., 2001). Another way to view locality is to recognize that the truth value of a modal formula can change in a model depending on the world of evaluation: if the latter changes, the context of accessible points may change as well. The property of globalness then becomes an 'unmodal' property, and has two corresponding conceptions: the satisfaction of a formula is global if points inaccessible from the current state are relevant to truth or falsity; or if its truth value is the same in a model for all worlds of evaluation. A well-known example of global formulas where these conceptions overlap is the *global modality* (Blackburn et al., 2001; ten Cate, 2004):

Global diamond E: $\mathsf{M}, w \vDash \mathsf{E}B$ iff $\mathsf{M}, v \vDash B$ for *some* state $v$ in $\mathsf{M}$

Global box A: $\mathsf{M}, w \vDash \mathsf{A}B$ iff $\mathsf{M}, v \vDash B$ for *all* states $v$ in $\mathsf{M}$

The idea that inaccessible states affect the truth value of a formula can be made precise by the notion of violating invariance results under disjoint unions. Note that this is an instance of BR, even though we only treated isomorphisms and bisimulations in Section 6.2 (namely, by taking as model equivalence not bisimulation or isomorphism, but disjoint union as in Definition 5.3.2.). Instead, world invariance as a measure of globalness is stricter than BR. We will show soon that BR is indeed implied by world invariance[4], and we take the latter as our conception of globalness. A border case that is considered local by this conception is $\forall a A$, that intuitively possesses some type of globalness. We will discuss this formula more in Section 6.3.2.

Besides straying from an intrinsically modal nature, global formulas may be seen as more semantic than local formulas. Local formulas, only taking into account accessible states, are simply blind to a certain part of a Kripke model. Global formulas, when their truth value depends on inaccessible states, can collect more information about the model — and by the property of world invariance, they lift information up to a state that the entire model finds itself in. That is, instead of forming truth relations with each world separately, these formulas provide a truth state for all worlds in a model at the same time.

[4]Combined with valuation dependence or contingency.

**Valuation independence.** The second is the property of valuation independence. Generally, a modal formula "is valid on a frame when it is globally true, no matter what valuation is used. This concept allows modal languages to be viewed as languages for describing frames" (Blackburn et al., 2001). This quote connects both globalness and valuation independence to semantic properties, but we here focus on valuation independence separately. The semantic nature of such formulas is then still best seen when first restricting to the modal language. Valuation independent modal formulas can be seen to describe the local frame structure of the world of evaluation $w$, as shown by the simple example $\Diamond\top$ ('$w$ has a successor'). However, they are still instances of formula types that we do not regard as semantically polluted, because operators like $\Diamond$ are primarily intended to capture valuation dependent statements, and they clearly do not satisfy BR. On the other hand, if an operator introduced in the proof-theoretic language can *only* convey valuation independent information, then we consider it semantically polluted. In case such a formula is translatable to the modal language, its semantic nature is strengthened by the fact that it will be describing the frame. There can also be valuation independent formulas that are untranslatable to the modal language. They can still describe frame structure (such as a formula that is true when the world of evaluation has exactly two successors), but they can also concern other properties of worlds (concerning, for instance, an assignment function of an extended model). A more general view on the semantic nature of valuation independence then says that they have a stronger connection to the world of evaluation than the modal language.

We can thus consider formula types satisfying one of these properties as carrying a different type of semantic pollution as formula types just satisfying the base requirement. Furthermore, formula types satisfying both globalness and valuation independence can be thought of as possessing the strongest form of semantic pollution. The gradations of satisfying the base requirement can still show differences in semantic pollution within these four new categories.

### 6.3.1 Defining the levels

The properties of globalness and valuation independence are made precise by the following definition.

**6.3.1 Definition (Modal semantic properties).** Let $C$ be an operator or metavariable of PL. The modal semantic properties are then defined with respect to PL.[5]

[5]This is a choice: they can also be defined with respect to L, for instance with our earlier motivation of capturing only the effects of $C$, and not interactions with other operators. However, for our operators, the results relative to the different available context languages remain the same. Additionally, as mentioned before, $\downarrow aA$ and $\forall aA$ simply act like $A$ when $A \in \mathsf{L}$, so in order to see some more interesting examples highlighting the workings of $\forall$ and $\downarrow$, it is insightful to let $A \in \mathsf{HL}$, and so to let

The notion of globalness gives rise to two variants concerning the *relation of $C$ to states*.

1. **Locality (LO).** There is an $A \in C(\mathsf{PL})$, a model PM and worlds $w_1$ and $w_2$ such that $\mathsf{PM}, w_1 \vDash A$ and $\mathsf{PM}, w_2 \nvDash A$.

2. **Globalness (GL).** For each model PM and for all $A \in C(\mathsf{PL})$ it is the case that either:

   (a) $\mathsf{PM}, w \vDash A$ for all $w \in W$ ($A$ is *globally true in PM*), or

   (b) $\mathsf{PM}, w \nvDash A$ for all $w \in W$ ($A$ is *globally false in PM*)

The second property, also with two variants, concerns the *relation of $C$ to the valuation.*

1. **Valuation dependence (VD).** There is an $A \in C(\mathsf{PL})$, a pointed frame $(\mathsf{PF}, w)$, and two models $\mathsf{PM}, \mathsf{PM}'$ extending PF such that $\mathsf{PM}, w \vDash A$ and $\mathsf{PM}', w \nvDash A$.

2. **Valuation independence (VI).** For each pointed frame $(\mathsf{PF}, w)$, and for all $A \in C(\mathsf{PL})$, it is the case that either:

   (a) $\mathsf{PM}, w \vDash A$ for all models PM over PF (we say $A$ is *$w$-valid on PF*), or

   (b) $\mathsf{PM}, w \nvDash A$ for all models PM over PF (we say $A$ is a *$w$-contradiction on PF*)

As these properties will be imposed on top of the base requirement, they only function to further classify formula types that are already untranslatable to the modal language. Thus, now we can define the following four levels of semantic pollution (see Figure 6.5).

**6.3.2 Definition (Levels of pollution).** Let $C$ be an operator or metavariable of PL, and let BR be defined with respect to context language CL, model equivalence $E$ and (if applicable) extended equivalence strength $S$. Then four levels of pollution are defined as follows.

1. $C$ satisfies *weak semantic pollution* if it satisfies $\mathsf{BR}_{\mathsf{CL},E,S}$ + LO + VD.[6]

2. $C$ satisfies *local semantic pollution* if it satisfies $\mathsf{BR}_{\mathsf{CL},E,S}$ + LO + VI.

3. $C$ satisfies *global semantic pollution* if it satisfies $\mathsf{BR}_{\mathsf{CL},E,S}$ + GL + VD.

4. $C$ satisfies *strong semantic pollution* if it satisfies $\mathsf{BR}_{\mathsf{CL},E,S}$ + GL + VI.

the context language generally be PL.

[6]Note that here, the only reason $C$ is semantically polluted is that it satisfies some variant of BR.

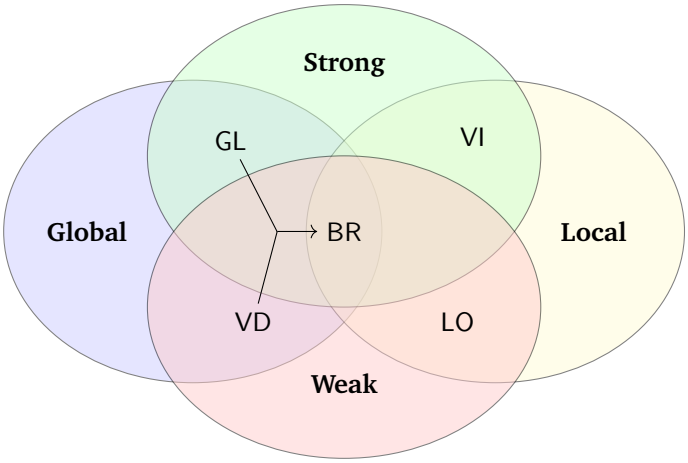



Figure 6.5: Levels of semantic pollution (the properties of GL together with VD imply the base requirement).

Blackburn et al. (2001) provide a simple proof that global diamond and global box are undefinable in the basic modal language. Similarly, we can provide an untranslatability result for the two properties of global semantic pollution (so that BR is actually a superfluous requirement here). Consider the definition of a disjoint union, followed by the theorem.

**6.3.3 Theorem.** *If an operator or metavariable $C$ of PL satisfies globalness and valuation dependence relative to CL, then it satisfies BR$[$CL$, \biguplus,$ FE$/$CE$]$.*[7]

*Proof.* Suppose that $C$ satisfies globalness and valuation dependence relative to CL. Then there is an instance $A \in C(\mathsf{CL})$ and two models PM and PM$'$, such that $A$ is globally true in PM and globally false in PM$'$. Now take an FE or CE- disjoint union PM $\biguplus$ PM$'$, which will be equivalent to both PM and PM$'$ (separately).[8] Since $A$ is invariant under worlds, one of the following holds:

1. PM $\biguplus$ PM$', w \vDash A$ for all $w \in W \cup W'$
2. PM $\biguplus$ PM$', w \nvDash A$ for all $w \in W \cup W'$

Suppose wlog that the first case holds. Then there exists a world $w \in W'$ such that

[7]Note that if we require that some $A \in C(\mathsf{PL})$ is contingent, then globalness by itself already implies BR. Without contingency, $\top$ satisfies globalness, and clearly does not satisfy BR (hence, we need valuation dependence).

[8]On the contrary, SCE-disjoint unions are only equivalent to one of PM or PM$'$, which prevents the proof from going through. As an example, consider the instance $@_a p$ of the global and valuation dependent $@_a(\mathsf{L})$. Under an SCE-disjoint union it will retain its truth or falsity, as $a$'s range in the disjoint union must be the same as in the equivalent model.

$\mathsf{PM}', w \not\models A$, and such that $\mathsf{PM} \biguplus \mathsf{PM}', w \models A$. Thus, $A$ violates invariance under disjoint unions. □

As this theorem gives the BR-result for disjoint unions, and disjoint unions are a specific type of bisimulation, the result is implied for bisimulations as well. However, it is not implied for isomorphisms. A counterexample is a global modality, for instance A$p$ or E$p$, which are global and valuation dependent, yet invariant under isomorphisms.

Now we provide a few remarks concerned with the difference between formula types $C(\mathsf{PL})$ and concrete instances $A \in C(\mathsf{PL})$ of a formula type. First, observe that global semantic pollution requires global matters to have full influence on the truth value of formulas for it to result in pollution. Suppose that a proof system introduces an atomic formula $A$ translatable to the specific formula $\forall a(a \wedge p)$, or similarly to $xRy \wedge p$. The parts $\forall a(a)$ and $xRy$ are global and violate invariance under (for instance) FE- and CE-disjoint unions (for these notions, just add disjoint unions to the regular model equivalences in Section 6.2). However, $p$ introduces not only valuation dependence but also locality into $A$, so that both versions of $A$ (as atomic formula types) are not globally semantically polluted. This means that covering up globalness with local 'camouflage' is considered to decrease semantic pollution, as this means that it is not the primary intent of a formula to convey just a global property. Analyzed at their main operator, our method still disects $\forall a(a \wedge p)$ and $xRy \wedge p$ into syntactically pure parts ($\wedge$ and $p$) and parts possessing semantic pollution ($\forall aA$, $a$, $xRy$).

As for the definition of local semantic pollution, note that it considers modal formulas that are local and valuation-independent (like $\Diamond\top$) to be syntactically pure. This is seemingly because of BR, which ensures that semantically polluted formulas are not translatable to the modal language. However, even without BR $\Diamond\top$ would not count as semantically polluted, as we measure pollution at the level of formula types. Clearly, $\Diamond(\mathsf{L})$ does not satisfy valuation independence. Still, for each modal instance of a local and valuation independent formula, a new primitive formula type $A$ can be added to the modal language that has exactly the truth condition of this instance (such as '$w$ has a successor'). Requiring BR then means that formula types $A$ that are translatable to concrete local, valuation independent modal formulas, are pure just like these modal formulas themselves — while without BR, such formula types $A$ *are* semantically polluted (even though their translations are not). Hence, our definition of local pollution, including BR, says that a formula is only locally semantically polluted if you cannot in principle replace it in a formal proof by an expression of the object language.

The way that the modal language can describe a Kripke frame (by being valuation independent) is thus considered acceptable, and a syntactically pure baseline. This includes specific descriptions of $R$-depth, as the modal operators describe exactly one $R$-step. The 'height formulas' in (ten Cate and Koudijs, 2022) show that

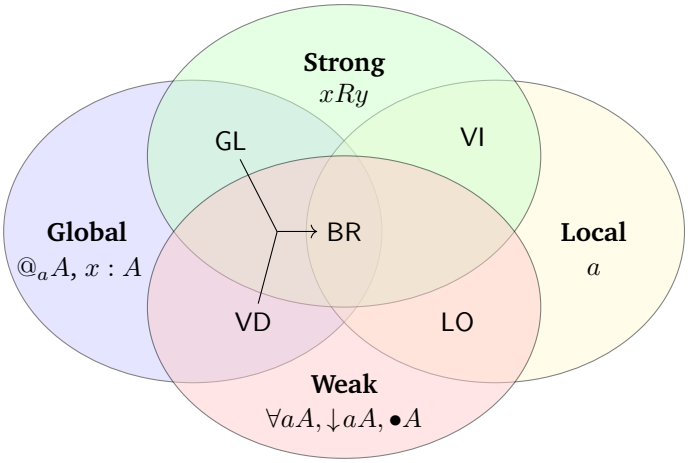



Figure 6.6: Levels of semantic pollution (the properties of GL together with VD imply the base requirement).

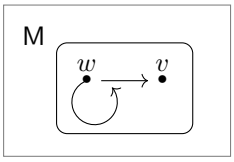



Figure 6.7: Example of locality.

a formula $\Box^{n+1}\bot \wedge \Diamond^n \top$ says that $w$ starts at least one $R$-path of length exactly $n$. The modal language cannot describe specific $R$-width, however, as a consequence of modalities describing 'at least one successor' or 'for all successors'. Specifying a precise number of successors thus provides more opportunities for introducing semantic pollution than specifying the length of an $R$-chain.

### 6.3.2 Results

Figure 6.6 shows the general division of the formula types in terms of the four levels of pollution. The results of locality and globalness are quite easily seen. Relative to their relevant proof-theoretic language, locality of $\bullet A$ should be obvious, as should globalness of $xRy$, $@_a A$ and $x : A$. Below are some examples of locality of the remaining HL-formulas.

**6.3.4 Example (Locality).** Consider model M in Example 6.7. Let $\tau(a) = w$. Then $\mathsf{M}, \tau, w \vDash \forall a(\Diamond a)$, as wherever $a$ is sent to, $w$ sees it. However, $\mathsf{M}, \tau, v \nvDash \forall a(\Diamond a)$, as $v$ does not see itself nor $w$. Additionally, $\mathsf{M}, \tau, w \vDash\!\downarrow a(\Diamond a)$, as $w$ sees itself,

while $\mathsf{M}, \tau, v \not\models \downarrow a(\Diamond a)$, as $v$ does not see itself.[9] And clearly, $\mathsf{M}, \tau, w \models a$, while $\mathsf{M}, \tau, v \not\models a$.

The results of valuation independence are even quicker to see. It is easy to tell that only $xRy$ and $a$ have a truth value that does not vary under changing propositional valuations.

Finally, as promised, some reflection on the hybrid operator $\forall aA$. It does not satisfy our criterion of globalness, but intuitively, its truth value *is* affected by inaccessible worlds ($A$'s truth is tested for $a$ sent by $\tau$ to all worlds in turn). To give body to this intuition, note that $\forall aA$ would satisfy a criterion of globalness as follows, a variant of violating invariance under disjoint unions. Given a model PM, add a separate Kripke model $(W, R, V)$ to it (essentially creating a disjoint union, but keeping $\tau$ of PM constant (in case of an extended model)). A formula type is then 'global' (i.e., able to be affected by inaccessible worlds) if its truth value can change under this model operation. Clearly, this is the case for $\forall a(\mathsf{HL}_{Q,V})$ (consider $\forall a(a)$), while it is not the case for $\downarrow a(\mathsf{HL}_Q)$ (as shown by an easy induction proof). This illustrates how $\downarrow aA$ is really just a specific subcase of $\forall aA$, sending $a$ to the current world, instead of all worlds in the domain. Thus, $\downarrow aA$ in some sense possess more locality than $\forall aA$, so that $\forall aA$ may deserve a higher level of semantic pollution.

However, note that the latter idea for globalness would, as $\tau$ remains static, not consider any labeled formula to be global, which is undesirable. A variant that would attribute semantic pollution to labeled formulas is violating invariance results under CE-, FE- or SCE-disjoint unions — but clearly, this is just another variant of BR, which satisfaction we already require. On top of that, recall that BR itself already captures a difference in pollution level between $\forall aA$ and $\downarrow aA$ (see Figure 6.4). Hence, after taking in the four levels of pollution defined here, more nuance within these levels can be found by looking at the levels of BR-satisfaction.

Finally, note the difference between $\forall aA$, $\downarrow aA$ on the one hand, and labels on the other: $\forall aA$ and $\downarrow aA$ are still *general* (it is unclear which world exactly $a$ refers to), while name variables in $@_a A$, $x : A$ and $xRy$ are always *specific* (they pinpoint particular points in a model). This corresponds intuitively to the idea that name variables in the latter operators are more semantically polluted than $\forall aA$ and $\downarrow aA$.

### 6.3.3 Four levels compared to the base requirement

To finish this section off, we highlight several observations that come out of the comparison of the four pollution levels to the results concerning BR of Section 6.2. First, $\bullet A$ *is a minimal example of semantic pollution.* It steadily has the lowest level of pollution, according to the four levels of pollution as well as BR. Within weak

[9] Locality of $\forall aA$ and $\downarrow aA$ is already clear from modal instances $\forall a(p)$, $\downarrow a(p)$, but hybrid local instances provide more insight.

semantic pollution, it may be considered to possess a lower amount of pollution than $\forall aA$ and $\downarrow aA$. On our approach, it thus possesses (in our context) the least number of properties required to be semantically polluted. For BR, it only violates invariance under bisimulations (consider generated submodels), but it is neither global nor valuation-independent.

Second, *BR brings nuance within weak semantic pollution and global semantic pollution.* Within weak semantic pollution, BR shows that $\forall aA$ possesses the highest level of pollution in this category, leaving $\downarrow aA$ in between $\forall aA$ and $\bullet A$. Furthermore, while $@_aA$ and $x : A$ are put in the same category of global pollution, BR shows that $@_aA$ is able to express more semantic pollution because of its stronger context language.

Third, *the local-global distinction tears apart $a$ and $xRy$.* While $a$ and $xRy$ seemed equal in their level of pollution with respect to BR, the four levels tear them apart. Their difference resides in that $a$ retains locality, while $xRy$ does not. In doing so, $xRy$ becomes the maximal example of pollution among our collection of formulas.

We conclude that the four levels of semantic pollution create a helpful overview of the differences between our formula types, while BR remains a useful tool to gain more refined insights into semantic pollution results. The results show that display calculi are very weakly polluted, hybrid calculi possess an intermediate level of semantic pollution (with three different variants), and labeled calculi have the strongest level of pollution (with two different variants).

## 6.4 Philosophical views on semantic pollution

The previous sections gave us a definition of the intuitive phenomenon of semantic pollution of modal proof systems. As shown in Section 5.4.1, the discussion in the literature concerning philosophical suitability of semantically polluted proof systems is ongoing. Instead of aiming to solve the matter in this paper, we will propose that our characterization of semantic pollution is neutral with respect to the debate on suitability for inferentialism; that it emphasizes importance of the distinction between implicit and explicit proof systems; and that it is compatible with less often voiced reasons for desiring syntactic purity.

### 6.4.1 Suitability of proof systems for inferentialism

The literature discusses the suitability of proof systems with extended proof-theoretic syntax for inferentialism, i.e. the idea that the meaning of logical connectives is established by their inference rules, instead of model-theoretic semantics (see for instance Schroeder-Heister (2024)). Most notably Read (2015); De Martin Polo (2024) have provided a philosophical defense of semantically polluted (in particular, labeled) calculi for inferentialism.

One view in this debate, as advocated by Read (2015); De Martin Polo (2024), says that properties such as harmony (and separability, and others) are decisive in determining suitability for inferentialism, no matter the proof-theoretic language used. Then, labeled calculi are acceptable, as "[t]he labeled rules [...] for $\Box$ and $\Diamond$ are harmonious, that is, the introduction rules encapsulate the whole meaning of the modal operators" (De Martin Polo, 2024). Read (2015) emphasizes that "[t]he semantics lies in the shape of the rules", and so relational atoms and labels "need not be thought of as having any meaning themselves".

A strand of more philosophically-oriented proof theorists argue that proof systems should suitably correspond to our inferential practice (in order to be suitable for inferentialism). This idea also comes in independently for those who use a proof system to faithfully formalize informal reasoning. Its application to inferentialism for instance emerges as Steinberger (2011)'s Principle of Answerability: "[a]dherence to inferentialism importantly constrains one's choice of proof-theoretic frameworks and thus requires one to reject Carnap's amoralism about logic: the inferentialist must remain faithful to our ordinary inferential practice". Not only semantically polluted calculi are subject to this view, but also syntactically pure proof-theoretic languages: for instance multiple-conclusion calculi (see Steinberger (2011); Restall (2005)), and hypersequent calculi (see e.g. Hjortland and Standefer (2018)). On this view, it is unclear where labeled calculi stand, although at first sight they seem a rather controversial idealization of modal reasoning in practice.[10] These arguments are burdened by the question of what an 'acceptable' idealization is: Dicher (2020) for instance claims that such a boundary is unhelpful, and any idealization should be acceptable.

Our characterization of semantic pollution in this paper is relatively neutral with respect to these aspects of inferentialism. Semantic pollution (in terms of satisfying the base requirement, valuation independence, or globalness) nor syntactic purity prevents or guarantees harmony. Concerning 'principles of answerability', our definition may at first sight be seen as a proposed 'border' for acceptable (semantically polluted) and unacceptable (syntactically pure) idealizations. Semantically polluted calculi analyze the use of a connective in a stronger language than the object language: such strong 'language contexts' are perhaps more prone to unacceptable idealization. However, we discourage such a strict view: strong languages are clearly not guaranteed to be separated from inferential practice (lots of natural logical languages differ in strength), while weaker languages can still be shaped artificially and lack correspondence to practice (just consider the debate on multiple-conclusion calculi (Steinberger, 2009)). In the end, we simply encourage inferentialists using syntax-rich proof systems to spell out the intended

[10] In terms of the possible world interpretation, there are at least certainly counterexamples. For instance: "Why are horses necessarily mammals? Not because every horse is a mammal in every possible world. But because the property of being a horse bears a special relationship to the property of being a mammal." (Warmke, 2016) As for a temporal interpretation of labels, Arthur Prior (see (Blackburn, 2006)) for a long time considered the use of labels (in hybrid logic) to promote an unnatural 'reasoner-external' perspective, as opposed to our internal experience of time.

interpretation of the syntax in the context of inferential practice.

### 6.4.2 The relevance of distinguishing explicit and implicit proof systems

A further justification for the use of labeled rules in inferentialism draws on the distinction between explicit and implicit rules. In the literature, labeled systems are considered to incorporate the semantics 'explicitly'. Explicit semantic elements are made precise by Poggiolesi (2010) as the idea that sequents containing them are untranslatable to the modal language. Instead, systems such as nested sequents or tree-hypersequents (that do have an interpretation into the logic), then import semantic elements implicitly.[11] Authors often consider the implicit incorporation of semantic elements to be syntactically pure (Poggiolesi, 2010; Brünnler, 2010), while (explicit) labeled calculi are generally considered semantically polluted.

De Martin Polo (2024) and Read (2015) argue that the difference between explicitness and implicitness is less big than it seems. They argue that there is no actual semantic difference between them when considered closely — hence, labels should not be considered more semantically polluted than implicit calculi. The argument in (De Martin Polo, 2024) goes:

> "Read notes that in tree-hypersequent calculi, the semantic content is still explicitly present, but is indicated by the symbols "/" and ";" instead of R. Similarly to Boretti, he argues that the tree-hypersequent rules for necessity only encode the semantic structure of modal formulas in an opaque and disguised manner, thus simply obscuring the semantic apparatus that is more evident in the notation of labeled sequents [...] even though the apparatus of Kripke semantics is presented differently in tree-hypersequent systems than in labeled calculi, it is still (i) explicitly displayed (although obscured in an unconventional notation) and (ii) plays a fundamental role."

While we remain neutral on the difference between implicit and explicit proof systems regarding suitability for inferentialism, we here maintain that this difference *is* relevant regarding the notion of semantic pollution. As claimed above, it is true that the labeled calculus and the nested or tree-hypersequent calculus both arrange modal formulas into a graph structure: but this fact by itself is not enough to claim that their relation to the Kripke semantics is the same. In this paper, we suggest that semantic pollution arises from the *way* that the graph structure is described. If it is described by a language that can express more about Kripke models than the modal language can, then 'too much' detail about the semantics enters the language, and we talk of pollution, and of an 'explicit' calculus. If, like

[11] As mentioned, this is similar to the distinction between 'external' and 'internal' proof systems, see also (Lyon et al., 2023).

nested and tree-hypersequent calculi, the graph structure is described by formulas that have an interpretation in the logic (and so a truth condition like that of a logical formula), then this graph structure has no particular 'semantic' nature at all — in any case, it is no more semantic than the logic itself. This means that we do not think that "in tree-hypersequent calculi, the semantic content is still explicitly present". Rather, the semantics is described *through* the object language. Of course, it may still be the case that proof-theoretic syntax which has a logical counterpart, has a different informal meaning (e.g. in inferential practice) than its logical translation. However, our point here is that the relation to the model theory of the proof-theoretic syntax and its logical counterpart (by their truth conditions) is in fact the same.

Saying that "/" and ";" display Kripke semantics at the level of tree-hypersequents, is like saying that $\Diamond$ and $\Box$ display the semantics at the level of the basic modal language. Should we thus consider $\Diamond$ and $\Box$ as semantically polluted? It seems clear that this is not so. Simply as a consequence of its model-theoretic truth conditions, *any* syntactic element displays the semantics to some extent. The interesting question for semantic pollution is where the boundary lies: what syntactic elements do we consider pure (surely, the logical object language) and what syntactic elements do we consider impure (our proposal is found in the previous sections).

This relates to our conceptual understanding of the different ways in which proof calculi can describe Kripke frames (see also Poggiolesi and Restall (2012)): labeled systems can explicitly and globally describe a Kripke frame, while display calculi use a local perspective while allowing perspective switches along $R$-connections (i.e., switches between actual worlds). By incorporating the forward as well as backward perspective, display calculi can describe a Kripke frame better than the modal language. The tree-hypersequent or nested systems, on the other hand, through their logical interpretation, only possess the (syntactically purest) local 'forward' perspective: everything is encoded through uses of $\Box$.

Note also that this counters the 'proof-theoretic' idea that 'notational variants' of proof systems should have an equal level of semantic pollution.[12] There exist proof-theoretic translations between labeled sequents, nested sequents, (tree-)hypersequents and display sequents (see e.g. Ciabattoni et al. (2021); Goré and Ramanayake (2014), and the hierarchy of translations defined in (Lyon et al., 2023)). Now consider for instance a labeled calculus that only allows labeled *tree* sequents (so that it is formally now 'just a notational variant' of a nested or tree-hypersequent calculus, in which all structures are already trees). Our proposal will still say that this labeled calculus is semantically polluted, whereas the nested and tree-hypersequent calculus is not: the labeled calculus still describes a tree using more expressive power than necessary. It uses relational atoms and labels that, by definition, can express more semantic information than the nested structure.

[12] See (French, 2019) for an analysis of when two logics can be said to be notational variants — note that it is unclear whether the intuitive use of notational variant here corresponds to this analysis, although this is not the place to further investigate this.

And perhaps more importantly: if made true model-theoretically, the labeled tree sequent will actually indicate a tree-form in the Kripke model, by indicating the specific worlds and relations. A nested sequent or tree-hypersequent has a tree structure within the sequent, but its model-theoretic truth at a world (interpreted in terms of disjunction and box) does not enforce this world to be arranged in a tree inside the model. That is, even though the different proof systems may describe the same graph arrangement in a sequent, the labeled calculus does this in a *semantically polluted way*. This shows that proof translations do not always preserve philosophical values.

Hence, on our approach, the distinction between explicit and implicit notions matters when defining semantic pollution: it provides a natural formal definition supporting a distinction that philosophers and proof theorists already make intuitively — and we thus conclude that this distinction cannot be abolished on the grounds that their semantic content is the same.[13]

### 6.4.3 Syntactic purity as an ideal of proof (systems)

Finally, we argue here that there exist clear cases where semantic pollution has philosophical harm, other than possible harm for inferentialism, and for closeness to inferential practice. First, motivations of aesthetics, or ideals of proof, are and have always been common in mathematical fields. Dawson (2006) presents an overview of reasons mathematicians have to reprove a theorem, among which "to employ reasoning that is simpler [...] than earlier proofs". Simplicity is a well-known ideal of proof, where we may distinguish between conceptual simplicity, and formal (computational) simplicity. In our analysis of proof systems, we can make a similar distinction. labeled calculi may conceptually provide a simple way of analyzing inference rules (for those familiar with Kripke semantics). However, the ideal of simplicity may also manifest itself in the search for a proof system in as small a language as possible. I.e., as also noted by Lyon (2021b, p. 112), we might be interested in finding out that there exists a satisfying proof system for modal logic that is restricted to the modal object language, without any 'brute force' or potential clutter from external syntax. No more than aesthetics (a desire for resource-minimality) is involved in this, and yet it is a common motivation in mathematics. From this point of view, semantic pollution is in fact undesirable. Incidentally, other reasons than aesthetics may still also apply: Lyon (2021b) notes that a tree structure of sequents (that nested sequents or tree-hypersequents guarantee, but labeled sequents do not) can be necessary for certain proof-search algorithms. In this case, the excess of syntax in semantically polluted systems can even provide too much freedom for technical applications.

[13] See Section 7.1 for some more general reflections on our choice of formalization for semantic pollution, and Example 7.3.1 for a refinement in the distinction between explicit and implicit proof systems, illustrating that not *all* untranslatable syntax should be considered to be semantically polluted.

Another ideal of proof (systems) may be, given a model-theoretic semantics for it, that the proof system and the model theory are sufficiently conceptually separated. This is possibly simply what Avron (1996) meant with his claim that a proof calculus should be 'independent' from any particular semantics: the fact that we consider proof theory to be an activity that is somehow separate from model theory. Similarly, soundness and completeness proofs should then show us a 'valuable' insight: instead of giving us two notational variants of the same conceptual approach to a logic, they should connect a model-theoretic perspective to a (sufficiently different) proof-theoretic perspective. If not, the 'proof system' can be regarded merely as a systematization of semantic thought. Although not much more than aesthetics seems to validate these preferences, in a bottom-up approach to philosophy of proof theory, they should be taken seriously.

A different motivation for discarding semantic proof systems, in particular labeled systems, has to do with *impartiality* with respect to the background logic. Given a desire for a model-theoretic semantics, and the natural interpretation of labels into Kripke semantics, there is arguably a sense in which labeled calculi favor a classicist world-view. Classicists who want to reason with modalities may be happy to accept labels as concerning time or possibilities. Intuitionists who wish to assign some interpretation to labels and relational atoms, may struggle to find a satisfying one: labeled rules would ask them to quantify over worlds (or times), and explicitly refer to states other than the actual one. This simply may not be acceptable for them, even though modal reasoning should be a perfectly acceptable activity for intuitionists. An interesting open question relating to the latter two points (which we leave to future research) is how a proof system is *formally* independent or impartial from a particular semantics. That is, perhaps there is an interesting way to say that the relation of labeled calculi to Kripke semantics is more 'necessary' formally, than its relation to other types of semantics.

## 6.5 Conclusion

We have presented a characterization of semantic pollution of proof systems in terms of four levels of pollution. Our measures suggest that the nature of modal syntax lies in what it can express about Kripke models, its local view of a model, and direct interaction with basic propositions. Instead, the properties of higher expressivity than the modal language, a global view of a model, and valuation independence suggest semantic influence in proof-theoretic syntax.

Our results show that the display calculus is only weakly semantically polluted (by $\bullet A$). The hybrid calculus, on the other hand, introduces formula types that are weakly semantically polluted ($\forall aA, \downarrow aA$), but also ones that possess global ($@_a A$) and local ($a$) semantic pollution. Finally, the labeled calculus introduces a globally polluted formula type ($x : A$) and the only strongly semantically polluted formula type ($xRy$). In line with intuitions throughout the literature, then, the labeled

calculus can be seen as possessing the highest level of semantic pollution (among the calculi that we studied). We concluded that the difference between explicit and implicit proof calculi is key in our characterization of semantic pollution, and that besides the virtues of polluted calculi, semantic pollution can just as well have technical and philosophical downsides.

We might also seek a more general analysis of semantic pollution as a distinction between syntactic proof systems and their semantics, with applications to all kinds of logics. Indeed, semantic pollution might occur in all logical areas where 'good' proof systems are hard to find. The measures of violating invariance results under model equivalences, and being valuation independent, seem relatively easily transferable to other logics. World invariance seems more tailored to a type of semantics with different 'points of evaluation', although extensions to intuitionistic logics seem natural. These quick considerations already give reason to think that certain proof systems for neighborhood semantics (similar to labeled systems) are semantically polluted (see (Dalmonte et al., 2018), and similarly (Negri, 2016)), as well as a proof system for intuitionistic predicate logic (see (Baaz and Iemhoff, 2008)).

Besides the extension of semantic pollution to other logics, future research could analyze semantic pollution of modal proof systems for different types of proof systems than the ones we chose to study; or they could provide additional conceptions of semantic pollution that come with different measures. These directions could all provide us with a more fundamental understanding of the difference between syntax and semantics. The next chapter will reflect some more on possible generalizations of this measure of semantic pollution.

## 6.6 Appendix

We here provide the various induction proofs referred to in the paper:

1. $\downarrow aA$ is invariant under *FE-isomorphisms* relative to $\mathsf{HL}_V$ (see Theorem 6.6.1 and 6.6.2).
2. $\forall aA$ is invariant under *FE-isomorphisms* relative to $\mathsf{HL}_V$ (see Theorem 6.6.1 and 6.6.3).
3. All formulas in $\mathsf{HL}$ are invariant under *SCE-isomorphisms* (see Theorem 6.6.4).
4. All formulas in $\mathsf{HL}_Q$ are invariant under *SCE-bisimulations* (see Theorem 6.6.5).

The first proof will set the stage for showing that $\downarrow aA$ is invariant under FE-isomorphisms relative to $\mathsf{HL}_V$. Note that we cannot show that any hybrid formula $A$ has this property, as the base case where $A$ is a nominal $a$ violates it. The reason we show the general Theorem 6.6.1 before Theorem 6.6.2 is that the case

of $\Box A$ requires $a$ to be mapped to an arbitrary $y$ (instead of already sending $a$ to the current world as $\downarrow aA$ does), in order to apply the induction hypothesis at $R$-reachable worlds. We use $\cong_{\mathsf{FE}}$ as a symbol for FE-isomorphism, and $\cong_{\mathsf{SCE}}$ for SCE-isomorphism.

**6.6.1 Theorem.** *Suppose* $\mathsf{M}, \tau, w \cong_{\mathsf{FE}} \mathsf{M}', \tau', w'$ *and* $\mathsf{M}, \tau, y \cong_{\mathsf{FE}} \mathsf{M}', \tau', y'$. *Then for all* $A \in \mathsf{HL}_V$: $\mathsf{M}, \tau_{[a \mapsto y]}, w \vDash A$ *iff* $\mathsf{M}', \tau'_{[a \mapsto y']}, w' \vDash A$.

*Proof.* The proof proceeds by induction on $A$, with an induction hypothesis for $B$ less complex than $A$. Assume $\mathsf{M}, \tau, w \cong_{\mathsf{FE}} \mathsf{M}', \tau', w'$, and $\mathsf{M}, \tau, y \cong_{\mathsf{FE}} \mathsf{M}', \tau', y'$. Consider the cases below (we omit the straightforward cases of conjunction and negation).

- **Base case.** Clearly, a proposition letter $p$ has this type of invariance under FE-isomorphisms. For nominals, suppose $\mathsf{M}, \tau_{[a \mapsto y]}, w \vDash a$. Then $\tau_{[a \mapsto y]}(a) = w$ (and $y = w$). Now consider $\tau'_{[a \mapsto y']}$. As $\mathsf{M}, \tau, y \cong_{\mathsf{FE}} \mathsf{M}', \tau', y'$, and $y = w$, it holds that $w \cong_{\mathsf{FE}} y'$. As $\mathsf{M}, \tau, w \cong_{\mathsf{FE}} \mathsf{M}', \tau', w'$, it also holds that $w' = y'$. Hence, $\tau'_{[a \mapsto y']}(a) = w'$, and $\mathsf{M}', \tau'_{[a \mapsto y']}, w' \vDash a$.

- **Box.** Suppose $\mathsf{M}, \tau_{[a \mapsto y]}, w \vDash \Box A$. We need to show that for all $v'$ such that $Rw'v'$, $\mathsf{M}', \tau'_{[a \mapsto y']}, v' \vDash A$. By the isomorphism, for every such $v'$ there is an isomorphic world $v$ such that $wRv$. By assumption, for these $v$, $\mathsf{M}, \tau_{[a \mapsto y]}, v \vDash A$. Then by the induction hypothesis, for each corresponding $v'$ it holds that $\mathsf{M}', \tau'_{[a \mapsto y']}, v' \vDash A$. Hence, $\mathsf{M}', \tau'_{[a \mapsto y']}, w' \vDash \Box A$.

- **Satisfaction.** Suppose $\mathsf{M}, \tau_{[a \mapsto y]}, w \vDash @_a A$. Then $\mathsf{M}, \tau_{[a \mapsto y]}, \tau_{[a \mapsto y]}(a) \vDash A$. By assumption, $\tau_{[a \mapsto y]}(a)$ (that is, $y$) is isomorphic to $y'$. Then by the induction hypothesis, $\mathsf{M}', \tau'_{[a \mapsto y']}, \tau'_{[a \mapsto y']}(a) \vDash A$. That is, $\mathsf{M}', \tau'_{[a \mapsto y']}, w' \vDash @_a A$.

- **For all.** Suppose $\mathsf{M}, \tau_{[a \mapsto y]}, w \vDash \forall a A$. Then $\mathsf{M}, (\tau_{[a \mapsto y]})_a, w \vDash A$ for all $(\tau_{[a \mapsto y]})_a$. Note that this simply equals $\mathsf{M}, \tau_{[a \mapsto v]}, w \vDash A$ for all $v \in W$. By the isomorphism, for each $v \in W$ we have a world $v' \in W'$ such that $\mathsf{M}, \tau, v \cong_{\mathsf{FE}} \mathsf{M}', \tau', v'$. By the induction hypothesis, and by surjectivity of the isomorphism, we have $\mathsf{M}', \tau'_{[a \mapsto v']}, w' \vDash A$ for each $v' \in W'$. In other words, it holds that $\mathsf{M}', \tau'_a, w' \vDash A$ for all $\tau'_a$. So, $\mathsf{M}', \tau', w' \vDash \forall a A$, and we can also rewrite this as $\mathsf{M}', \tau'_{[a \mapsto y']}, w' \vDash \forall a A$ (as the assignment of $a$ does not matter).

- **Down-arrow.** Suppose $\mathsf{M}, \tau_{[a \mapsto y]}, w \vDash \downarrow aA$. Then $\mathsf{M}, (\tau_{[a \mapsto y]})_{[a \mapsto w]}, w \vDash A$. We can rewrite this as $\mathsf{M}, \tau_{[a \mapsto w]}, w \vDash A$. Now we apply the induction hypothesis to obtain $\mathsf{M}', \tau'_{[a \mapsto w']}, w' \vDash A$. Again, we can write this as $\mathsf{M}', (\tau'_{[a \mapsto y']})_{[a \mapsto w']}, w' \vDash A$, so that we obtain $\mathsf{M}', \tau'_{[a \mapsto y']}, w' \vDash \downarrow aA$.

□

**6.6.2 Theorem.** $\downarrow aA$ *is invariant under FE-isomorphisms relative to* $\mathsf{HL}_V$.

*Proof.* Take Theorem 6.6.1 and consider the instance where $y = w$ and $y' = w'$. The theorem then says that for isomorphic states $w$ and $w'$, $\mathsf{M}, \tau, w \models\!\downarrow aA$ iff $\mathsf{M}', \tau', w' \models\!\downarrow aA$. □

**6.6.3 Theorem.** *$\forall aA$ is invariant under FE-isomorphisms relative to $\mathsf{HL}_V$.*

*Proof.* We simply need to prove that for all models $\mathsf{M}, \mathsf{M}'$, assignment functions $\tau, \tau'$, worlds $w, w'$ and variants $\tau_a$ and $\tau'_a$, if $\mathsf{M}, \tau, w \cong_{\mathsf{FE}} \mathsf{M}', \tau', w'$ then

$$\mathsf{M}, \tau_a, w \models A \text{ for all } \tau_a \text{ iff } \mathsf{M}', \tau'_a, w' \models A \text{ for all } \tau'_a$$

So suppose $\mathsf{M}, \tau, w \cong_{\mathsf{FE}} \mathsf{M}', \tau', w'$, and that for some $A \in \mathsf{HL}_V$, $\mathsf{M}, \tau_a, w \models A$ for all $\tau_a$. This equals $\mathsf{M}, \tau_{[a \mapsto v]}, w \models A$ for all $v \in W$. By the isomorphism, for each $v \in W$ there is a world $v' \in W'$ such that $\mathsf{M}, \tau, v \cong_{\mathsf{FE}} \mathsf{M}', \tau', v'$. By Theorem 6.6.1, and by surjectivity of the isomorphism, we obtain $\mathsf{M}', \tau'_{[a \mapsto v']}, w' \models A$ for each $v' \in W'$. Thus, $\mathsf{M}', \tau'_a, w' \models A$ for all $\tau'_a$. □

**6.6.4 Theorem.** *All formulas $A \in \mathsf{HL}$ are invariant under SCE-isomorphisms.*

*Proof.* The proof is by induction on $A$, with an induction hypothesis for formulas $B$ less complex than $A$. We only treat the cases of nominals and of the hybrid operators (as the modal cases are straightforward). Suppose $\mathsf{M}, \tau, w \cong_{\mathsf{SCE}} \mathsf{M}', \tau', w'$.

- **Base case.** Suppose $\mathsf{M}, \tau, w \models a$, so that $\tau(a) = w$. By the requirements of SCE-isomorphisms, equivalent worlds satisfy the same nominals, and so $\tau'(a) = w'$, and $\mathsf{M}', \tau', w' \models a$.

- **Satisfaction.** Suppose $\mathsf{M}, \tau, w \models @_a A$. Then $\mathsf{M}, \tau, \tau(a) \models A$. By the induction hypothesis, and the fact that $\mathsf{M}, \tau, \tau(a) \cong_{\mathsf{SCE}} \mathsf{M}', \tau', \tau'(a)$, $\mathsf{M}', \tau', \tau'(a) \models A$. Hence, $\mathsf{M}', \tau', w' \models @_a A$.

- **For all.** Suppose $\mathsf{M}, \tau, w \models \forall aA$, so that $\mathsf{M}, \tau_{[a \mapsto v]}, w \models A$ for all $v \in W$. We have to show that $\mathsf{M}', \tau'_{[a \mapsto v']}, w \models A$ for all $v' \in W'$. By the isomorphism, for each $v' \in W'$ we have a world $v \in W$ such that $\mathsf{M}, \tau, v \cong_{\mathsf{SCE}} \mathsf{M}', \tau', v'$. Specifically, for each such pair of worlds, $\mathsf{M}, \tau_{[a \mapsto v]}, w \cong_{\mathsf{SCE}} \mathsf{M}', \tau'_{[a \mapsto v']}, w'$ holds: changing $a$'s assignment does not break the SCE-isomorphism, as $a$ is still sent to equivalent worlds. By the induction hypothesis, and by surjectivity of the isomorphism, we have $\mathsf{M}', \tau'_{[a \mapsto v']}, w' \models A$ for each $v' \in W'$. Hence, $\mathsf{M}', \tau', w' \models \forall aA$.

- **Down-arrow.** Suppose $\mathsf{M}, \tau, w \models\!\downarrow aA$, so $\mathsf{M}, \tau_{[a \mapsto w]}, w \models A$. Then, we also have a strong isomorphism $\mathsf{M}, \tau_{[a \mapsto w]}, w \cong_{\mathsf{SCE}} \mathsf{M}', \tau'_{[a \mapsto w']}, w'$, as $a$ is still sent to equivalent worlds. By the induction hypothesis, $\mathsf{M}', \tau'_{[a \mapsto w']}, w' \models A$. Hence, $\mathsf{M}', \tau', w' \models\!\downarrow aA$.

□

**6.6.5 Theorem.** *All formulas $A \in \mathsf{HL}_Q$ are invariant under SCE-bisimulations.*

*Proof.* This proof proceeds in the same way as that of Theorem 6.6.4. Note that the base case works even for bisimulations, as SCE-bisimulations still need equivalent worlds to satisfy the same nominals. We skip the $\forall a A$ case (which does not work anymore because of possible change in model cardinality over the equivalence), but the cases of nominals, $@_a A$ and $\downarrow a A$ go through as before. □

# 7

# Reflections on ideals and formalization

We have traversed the landscape of ideals of proof and proof systems, and made several stops for case studies. It is now time to take a broader perspective, and to comment on possible connections between these case studies, and various potential generalizations. This will illustrate the characteristics of the general strategy that underlies our approach to these case studies, while our pluralist attitude will still recognize that there are a myriad of ways to formalize ideals of proof.

In particular, this chapter is divided into three parts. First, we will more closely consider the formalization of ontological purity compared to the formalization of semantic pollution in Section 7.1, to see where their conceptual as well as technical similarities lie and cease to exist. We then turn to possible generalizations of the measures of ontological purity (for formal proofs) and semantic pollution in Section 7.2, specifically focusing on different ingredients of a proof system, different types of proof systems and different background logics. Third, Section 7.3 aims to connect our formalizations to the general considerations on formalization provided in Chapter 1, specifying several virtues and dangers of our case studies, and reflecting on formalization of ideals of proof generally. The majority of this chapter should be considered as a preliminary broader understanding of the work done in the case studies, and as providing wide-ranging suggestions for further research.

## 7.1 Two formalizations compared

We may observe that the approaches to formalizing ontological purity and semantic pollution of formal proofs display several striking similarities, but also relevant differences. We discuss both patterns here, in order to better understand why they were chosen, and to extract a possible core strategy to formalizing proof ideals,

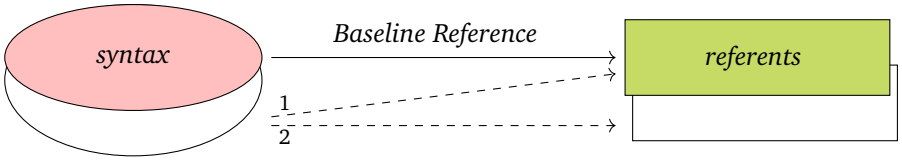



Figure 7.1: The reference of syntax (and extensions) to referents (and extensions).

that might benefit other ideals of proof as well. First, we elaborate on how the two formalizations should be understood as relating to each other conceptually, followed by a discussion of their respective use of formal tools.

### 7.1.1 The reference and the referents

A main similarity between the formalization of ontological purity and syntactic purity for formal proofs is that they both take a baseline relation of syntax *referring* to certain *referents* (see Figure 7.1), where this baseline represents an 'acceptable reference' satisfying the ideal of proof. In the previous chapters, our task lay in formalizing what this baseline reference amounts to, but also (as visualized by the dashed arrows in the figure) in specifying how new (extended) syntax can have an acceptable reference relation (compatible with the baseline reference), as well as what extensions of the referents yield an acceptable reference relation. Instead, syntax is considered to satisfy ontological impurity and semantic pollution if it goes beyond the accepted reference relation, i.e., to be incompatible with the baseline reference (possibly in different ways).

In the studies of ontological purity and semantic pollution, the baseline reference relation concerns the relation of syntax to a formal structure — more specifically, what pinpoints the referent is the interpretation, or truth condition, of a syntactic element in a(n) (intended) model of a mathematical or logical theory.[1] This means that the reference relation is taken to exist independently of a prover's epistemological state. Although the initial choice of the baseline reference contains a certain arbitrariness (in the selection of an ontology, and the selection of a type of logical semantics), suiting the informality of an ideal of proof, once settled it fixes the initial formalization of the ideal, and allows it to be refined or extended by (suitable) formal tools from that point on. We will compare the use of several related formal ingredients for purity and semantic pollution in the next sections. First, we stress that although both ideals start from a baseline reference of syntax to certain referents, this reference relation plays a different conceptual role in

[1]Recall that for ontological purity, this relation is at first *informal*, and if the intended standard model of a theory is adequate, it can be taken as a formal model-theoretic relation. Although a first-order model can always be seen as a formal structure only, neutral on its ontological interpretation (i.e., what types of mathematical objects are in its domain), assuming the adequacy of a standard model, we will at times use the term 'model' synonymously with 'ontology' in this chapter.

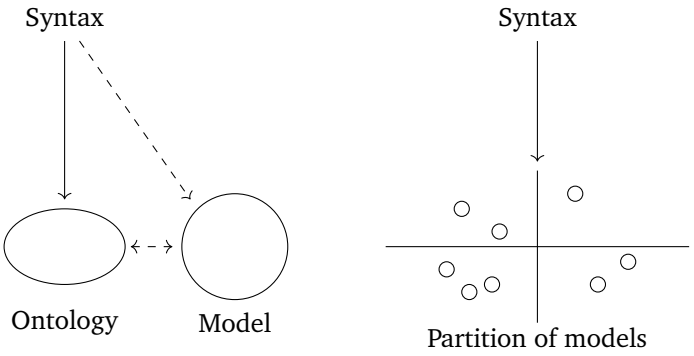



Figure 7.2: Visualization of the baseline reference relation for ontological purity (left) and syntactic purity (right).

determining whether purity or pollution holds.

In short, the difference is that ontological purity concerns the *referents*, while semantic pollution concerns the *reference*. This can be understood as follows. For ontological purity, the baseline relation consists of the indication of an ontology by a signature of certain primitives and a set of axioms in terms of them. New syntax can only be fully pure if it refers to exactly the right ontology, and nothing else: how this is done, is in principle irrelevant. As we have seen, one way for new syntax to be guaranteed to refer to the right ontology, is to be an abbreviation of a 'baseline syntax formula' (as in definitional extensions of the context theory); but other ways may exist. In order for new syntax to be secondarily pure, what is important is that they refer to surrogates of the right ontology. We found a way to ensure this by requiring that new syntax comes from a theory that interprets the context theory. Such syntactic elements then have 'good uses' (when they refer to surrogates) or 'bad uses' (when they do not). In short, the referent is the main indicator of purity.

Compare this to semantic pollution, where the baseline reference relation consists of the reference of a syntactic language to *all* models in a class that it is sound and complete with (see Figure 7.2). The baseline relation tells us how this language partitions the models into equivalence classes. This already suggests that it is not the particular referent that guarantees syntactic purity. Instead, in order to check whether newly introduced syntax is compatible with the baseline relation, then, we check whether it makes the same distinctions between the models. Note also that the occurrence of semantic pollution does not require any change of the model ingredients itself; nor is any particular ingredient of the model 'impure' from the beginning. Proof-theoretic syntax will describe the same referents as the object language, while the baseline reference relation sets the standard for *how* syntax is supposed to describe it.

**To simulate or not to simulate the referents**

Further observations should clarify the different ways in which ontological purity and syntactic purity accept simulations (or surrogates) of the baseline referents. First, observe the different levels on which model equivalences are used. For ontological purity, only one ontology is included in the baseline reference relation; and any extension of this relation to other referents concerns restricted ontologies of non-context theories. That is, even though the context theory will have many other models besides the one representing its ontology, these will not participate in an accepted reference relation, as we cannot guarantee their purity. Consider the many non-standard models of PA, for instance, containing non-standard numbers on top of all standard natural numbers — we certainly do not want to enforce that the intended 'pure' content of an arithmetical theorem contains such non-standard numbers.[2]

By contrast, in the case of semantic pollution, model equivalences are used to create a partition of a class of models, and so are already built into the baseline reference relation. We are less interested in using model equivalences in order to extend the baseline reference relation, as for ontological purity. That is, we have not used restrictions of models of other logical theories (or other types of semantics of the same logic, such as neighborhood, topological or algebraic semantics), to simulate the original syntactic purity result. Let us say a bit more on why our formalization does not allow any such 'secondary' syntactic purity result (elaborating on the remarks in Section 6.4.2). Generally, while our choice of formalization for ontological purity can be seen as one slowly widening the boundaries of a traditional interpretation of purity; the formalization of syntactic purity is better seen as a first, fully-fledged characterization of that notion. In that role, it aims to do justice to the unique status of labeled syntax as polluted, and as we will argue now, to a pre-theoretic understanding of syntactic purity as relative to a *fixed* class of models, and as fundamentally motivated by syntactic simplicity.

More specifically, one can understand syntactic purity as a 'semantics-relative' property. For ontological purity, we argued that a mathematical structure can in principle be instantiated by any mathematical objects, and so by any (powerful enough) type of semantics. But a partition of models is a view *on* a particular semantics. When replaced by a different semantics, our current position is that syntactic purity needs re-evaluation, because a 'view on a semantics' is only an interesting property relative to that particular semantics. It seems too fleeting a property to be suitable for simulation. In case one is interested in more tolerant (secondary purity-like) views, given a partition of models that is disconnected from its semantics, the question needs to be answered whether anything intuitively meaningful remains that we consider relevant for syntactic purity, and that can be

[2]Even models that are (mathematically) very similar to the 'pure' ontology might not suffice: consider a model containing just the even numbers, where $S$ is interpreted as $+2$, and so on — this might be a model of PA, but it is simply not the right ontology if we are looking for $\mathbb{N}$.

simulated through a translation. Consider an attempt: we might use a translation from an accepted proof-theoretic language to a semantically polluted language, and we restrict the uses of the polluted language to (more or less) translated expressions. These translations would originally be syntactic elements that equate the 'right' models. Hence, perhaps now, even though the polluted language makes a more fine-grained partition of models, the restricted use of polluted expressions might be said to 'simulate' the more coarse-grained partition. It remains unclear, however, in which exact sense translated syntactically pure syntax would approximate the truth conditions (embodying the perspective on the semantics) of the baseline pure syntax; and whether there is any interesting philosophical view on the nature of truth conditions that can motivate such a simulation. We leave such intriguing attempts to future studies.

A more fundamental reason for our current formalization choices may be found in the initial motivations for syntactic purity and ontological purity. Syntactic purity is an ideal of proof systems directly, whereas ontological purity originates from an ideal of *informal* proof. The intuitions of syntactic purity, then, (according to our current conception) concerns a certain *simplicity* of the base elements of the syntax (relating to the ideal of syntactic parsimony, see Chapter 1 and Section 6.4.3). This is something absolute, that translations are not guaranteed to preserve: just like interpretation translations, proof-theoretic translations can take a language with simple (or natural) base elements to a language with complex (or artificial) base elements. Then we might restrict the complex language to just translations of the simple language, but to 'simulate simplicity' is arguably not enough if we use too heavy machinery in doing this. Thus, we believe that a more absolute characterization of semantic pollution portrays intuitions surrounding syntactic simplicity better. For ontological purity, the ideal of informal proof does not place any immediate intuitive requirements on the 'simplicity' of the use of syntax; we just need to be convinced that the selected syntactic counterparts succeed in referring to the right ontology.

For now, these reflections should clarify the most important conceptual similarities and differences between our formalizations of of ontological purity and semantic pollution. A main difference concerns the baseline reference relation, where ontological purity focuses on the referents, while syntactic purity focuses on the reference. In the spirit of Chapter 1, however, we repeat that one may choose to focus on different pre-theoretic aspects of these ideals, so as to end up with different formalizations (that possibly share more of their conceptual underpinning regarding the reference relation).

### 7.1.2 Similarity of formal ingredients

We now observe that there are two main similarities between the use of formal tools for ontological purity as well as semantic pollution: the similarity between

definitional extensions and the formula interpretation; and the similarity between the interpretation translation and Kripke model equivalences.

For the first, consider the formal tools used for a conservative syntactic extension of the baseline reference relation (see arrow 1 in Figure 7.1), by which we mean that the extension preserves full ontological and full syntactic purity. In case of ontological purity, this concerns the extension of the context theory to its definitional extensions. In case of syntactic purity, this concerns the extension of a logical language, to proof-theoretic languages that have a formula interpretation back into the logic. Intuitively, these extensions are highly similar. A definitional extension obtained by adding to $\mathsf{T}$ an explicit definition of symbol $A$, can be seen to be the same as a proof-theoretic language extending a logical language $\mathsf{L}$ by syntax that has a formula interpretation (a sequent $\Gamma \Rightarrow \Delta$, for instance, can be seen as an 'abbreviation' of $\bigwedge \Gamma \rightarrow \bigvee \Delta$). Of course, the function of the extension is different: proof-theoretic syntax is not added to an object language along with an explicit definition of its formula interpretation — it is immediately brought into practice by incorporating it in logical axioms and rules of inference. We also repeat in the spirit of the previous section that the reasons why the extensions preserve full ontological and full syntactic purity, respectively, are different. In the case of ontological purity, it is of importance that the new symbol refers to the right ontological elements, whereas in the case of syntactic purity, we care that new symbols make the same distinctions between the models.

The second similarity concerns the extension of the ontological and syntactic purity result under 'equivalent' versions of the referents. Recall that this is included in the baseline reference relation for syntactic purity, and holds relative to the same syntax, defined by particular model equivalences. It corresponds to arrow 2 in Figure 7.1 for ontological purity, which holds relative to different syntax, and is induced by the syntactic interpretation translation. We might wonder whether the surrogates induced by interpretations satisfy particular model equivalences, just like the Kripke model equivalences we used. While again, the specifics should be elaborated on elsewhere, we here observe that surrogate ontologies seen as *internal models*, indeed seem to satisfy first-order model equivalences. The interpretation translation induces internal models that are created from models of the interpreting theory; and these are models of the context theory. For a full definition of internal models induced by the interpretation translation, see (Visser, 1997) (for a conceptual description, see also the end of Section 4.3.1).

Since such internal models will interpret the same signature as the standard model of the context theory, they can be related to the standard model by the usual first-order model equivalences. Arguably, due to their high similarity to the standard model (they will not contain any more objects or operations than the standard model), they will satisfy strong model equivalences such as isomorphisms (which are bijective homomorphisms, that in turn are structure-preserving functions between models, see e.g. Hodges (1993)). Hence, if we consider the ontology as the intended standard model of the context theory, the induced inter-

nal models are equivalent to the ontology. In a sense, then, the baseline reference relation is extended (through a secondary level of purity) to 'one structural equivalence class' of models, including the ontology and its induced surrogates; and satisfaction of the syntax of the context theory is invariant under these equivalences. We may interpret this as similar to the workings of Kripke model equivalences for syntactic purity, even though the level of syntactic purity stays the same under (and is defined by) these model equivalences, while the level of ontological purity does decrease relative to induced internal models.

Finally, conceptually, we might even see light similarities between a structuralist view on mathematical ontologies, and a local view on Kripke models, which are both *limited* views of the respective formal structures: in both cases, the language 'forgets' certain parts of the models. A surrogate ontology is captured by the 'good' formulas (as in Definition 4.3.1), and the domain formula $\delta$ makes sure that a part of the original ontology is cut away, simply forgotten. With respect to the remaining elements, the equivalence holds. Kripke model equivalences may also discard part of the model as not taking part in the equivalence (although isomorphisms do not). Both phenomena can be taken as a focus on 'core' information of the model, and leaving out unnecessary details. A final soft similarity may be seen between the language restriction for secondary purity, and the hybrid language restrictions to small formula types. In both cases, the (ontological or syntactic) purity result relative to such restrictions is lower than it would have been relative to the context theory and full HL, respectively. Note, however, that the nature of the language restriction is different in both cases: in small formula types, the restriction excludes specific operators entirely from the language; for secondary purity, the good formulas exclude certain *uses* of operators. Additionally, language restrictions in the hybrid language serve to consider the effects an operator more independently, or objectively; for secondary purity, the power of operators is reduced to *limit* their objective power, and simulate something else. We intend these last remarks to be taken lightly, and more detailed analysis of such surface similarities is necessary; however, they illustrate a broad common ground for the formalization of ontological and syntactic purity.

**7.1.1 Remark.** To close the comparison between ontological purity and semantic pollution, we remark that a major difference seems to be that semantic pollution is measured relative to the *logical* language, whereas ontological purity is measured with respect to the *non-logical* (mathematical) language. We might perhaps wonder whether the logical language might also induce breaches of ontological purity; and whether the non-logical language might also induce breaches of syntactic purity. If we take each ideal as strictly corresponding to just one of the formal categories of language (logical or non-logical), then this is by definition not possible. However, on more tolerant views this might be interesting to investigate more closely. For instance, on such views perhaps set-theoretic elements of the proof-theoretic syntax introduced by certain 'neighborhood' proof systems

(see (Negri, 2016)) might be able to affect the ontology of a formal theorem, by letting it include a domain of sets (even though the non-logical language might be, say, an arithmetical language). Here, there might be a case for saying that the proof-theoretic language creates ontological impurity. This is a specific example of the general idea that we might add any mathematical symbol (likely inspired by which mathematical structure makes up a logical semantics) to the proof-theoretic language. Still, our measures are currently separated by the distinction between the logical and non-logical language, and are suitable for their respective ideals as long as this classification is considered satisfactory.

### 7.1.3 Conclusions

Our comparison of ontological purity and semantic pollution illustrates the idea that a theorem can be seen as possessing a 'natural context' of an ontology and a structure (that encourages simulation of the ontology), as well as a partition of logical models (that is less suitable for simulation). Discrepancies of the proof concerning this context can cause violations of ideals of proof.

Given that two ideals of proof have suited a type of formalization including a baseline reference relation as in Figure 7.1, we might wonder whether other ideals of proof could also benefit from such a starting point. A candidate might be Pincock (2015)'s *abstract explanation*, that we have seen in Chapter 2. In short, this concerns the relation between mathematical objects mentioned by the theorem, and abstractions of these entities occurring in the theorem (and the existence of a biconditional between the mathematical objects and their abstractions). We might start to think about a formalization of such an ideal by considering our strategy for ontological purity, where we also made use of syntax referring to certain mathematical elements, and where a mathematical *structure* embodied an increase in abstraction. However, a foreseen difficulty for Pincock explanation is the fact that it is not obvious how to formalize abstraction differences relative to formal proofs. Pincock's model requires us to find a way to compare the level of abstraction of objects mentioned by the theorem, with those occurring in the proof. Like with ontological purity, we might associate the theorem with one formal theory, and the proof with another, and aim to establish an abstraction difference between these theories. However, as we have seen, interpretability and other notions of equivalence for first-order theories quickly equate entirely different theories with each other, and find a similar level of expressive detail in two theories that intuitively may possess different levels of abstraction. The power of syntax in first-order theories is simply often comparable, and we may not easily find a way in which multiple concrete objects instantiate one abstract object, the way Pincock intends the abstraction relation. However, interesting analyses may surely present themselves.

## 7.2 Potential generalizations of two formalizations

Our formalizations of ontological purity (of formal proofs) and semantic pollution are designed for specific proof systems with an established grammar, set of axioms and inference rules. Naturally, we may wonder what happens to these measures if we generalize them to proof systems that have a different balance between grammar, axioms and inference rules; or a similar balance, but different instances of these categories. Many things could be said on this topic; but we will restrict ourselves to some main expectations and suggestions.

We consider three types of generalizations. The first concerns the *level of analysis* of the measures with respect to the ingredients of a proof system. We may distinguish the levels of measures acting on the grammar; the axioms; the inference rules, and entire formal derivations. These levels put different emphases on the ideals of proof, and on different aspects of their role in reasoning. The second generalization concerns the *type of proof system*. We formulated our measures in terms of natural deduction (for purity), and variants of sequent calculi (for semantic pollution). For both ideals, it is helpful to see if there any interesting observations to make if we consider different types of proof systems. The final generalization concerns the behaviour of measures relative to changes in the *background logic*, which were also fixed in our case studies. We concerned ourselves with classical first-order logic, and classical modal logic. We might look at how these measures work for higher- or lower-order logics, as well as intuitionistic logics. Within the same category, we will consider changes of logical semantics: this is mainly relevant for semantic pollution.

### 7.2.1 Level of analysis

The formalization of ontological purity focuses on a proof system as an entire entity. That is, the formal measures already act on multiple ingredients of a proof system, and so generalizations of measures regarding their level of analysis are not so abundant. Full purity for formal proofs essentially relies only on the (grammar and) axioms of the context theory. However, logical inference rules are assumed to be purity-preserving, i.e., they cannot bring us from one ontology to another. In theory, of course, relating to the boundary between logic and mathematics — rules may be devised that do act on the mathematical ontology in an extraneous way. If they do, then additional measures are necessary to enforce full purity initially. For now, this just means that the requirement for inference rules to be 'logical' should be kept under scrutiny.

Secondary purity, on the other hand, relies on the axioms of the (interpreted) context theory, a grammar restriction of the interpreting theory, and restricted inference rules (that are the only allowed rules after the occurrence of 'pure formulas'). Thus, simultaneous requirements are given for acceptable axioms, rules,

and grammar: all ingredients work together to make a formal proof pure, and no single ingredient is enough to guarantee secondary purity. Hence, the formalization of secondary purity is comprehensive regarding the ingredients of a proof system.

Some more interesting observations can be made for semantic pollution. We made sense of semantic pollution completely by posing requirements on the proof-theoretic *grammar*. We can naturally wonder what semantic pollution looks like when it is defined specifically for different levels of analysis. Here we provide some brief thoughts on semantic pollution for inference rules. We first observe that extending the 'grammar measures' to work for inference rules seems too quick. For instance, one might want to assign syntactic purity to 'rules with a formula interpretation', but this seems to require turning rules into axioms, which in many settings (take the modal Necessitation rule, also discussed below) already cannot be done. Furthermore, in order to establish that a rule violates invariance under model equivalences, we would need truth conditions for rules; but the interesting part of truth conditions will be given by the grammar of its premises and conclusions. Hence, no natural way of saying that an inference rule (specifically) satisfies semantic pollution seems to arise. Let us consider two examples to gain some more insight into the possible behavior of semantic pollution for inference rules.

**Necessitation.** Consider the modal *necessitation rule*:

$$\frac{A}{\Box A}$$

It is widely known that this rule is not interchangeable with an axiom of the form $A \to \Box A$, due to the rule acting globally on a Kripke model, and the formulas inside the implication acting locally on it. This shows exactly that the rule is not translatable to the modal language (note that it thereby avoids the 'formula interpretation' guarantee for syntactic purity). Now suppose we take as a baseline reference relation the modal language with respect to a class of Kripke models that includes both normal and non-normal models (the latter can for instance be models that include 'non-normal worlds' that do not satisfy any boxed formula). The language will partition this class into equivalence classes which will sometimes reduce a normal and a non-normal model together (when non-normal worlds are not related to other worlds by the equivalence). Now suppose we introduce the Necessitation rule on top of this baseline. There will not be worlds that locally violate Necessitation, and so it cannot distinguish more models this way. Globally, however, the Necessitation rule will hold for all normal models, but will be violated for all non-normal models. This might happen within a previously defined equivalence class of the partition, and so the rule may satisfy a global version of BR.

We may also emphasize with this example that the class of models selected for the baseline reference relation affects the pollution result (already for our grammatic version of BR). If we consider a class of models that are all normal, comparing Necessitation to the reference relation will not create any semantic pollution, as the rule is valid everywhere. An example of this phenomenon for our original grammatic requirements is for instance a specific instance of the relational atom $xRx$, its truth condition expressing reflexivity of a specific world. While this quickly violates invariance under (some) model equivalences if we consider a bisimulation between a reflexive world, and two worlds seeing each only each other — it no longer does so if our base reference relation holds with respect to a class of reflexive Kripke models only. This emphasizes that a big enough 'sample' of models, allowing *contingency* of newly introduced proof-theoretic ingredients, is best for accurately analyzing semantic pollution (note that the original measures of Definition 6.3.2 allow non-contingencies in the strongest form of semantic pollution).

**The $\omega$-rule.** Consider next the usual variant of the $\omega$-rule (omega rule).

$$\frac{\varphi(0) \quad \varphi(1) \quad \varphi(2) \quad \dots}{\forall n \varphi(n)}$$

Some philosophical controversies surround this rule — for one, this concerns its relation to inferential practice. Although most people seem to think the $\omega$-rule is intuitive, and "instances of the omega-rule are at least informally valid" (McGee, cited in (Murzi, 2014)), it is also thought to be separated from inferential practice. This is because the rule works from infinitely many premises, and "it stretches the concept of a rule to encompass rules nobody is ever able to apply" (Peregrin, 2020). Contradictory sounds are also heard, claiming that "human beings do sometimes employ infinite rules of inference in their reasoning" (Brîncuş, 2024) (see also (Warren, 2021) for an elaborate philosophical argument on accepting infinitary reasoning). These arguments raise questions about the use of the $\omega$-rule for inferentialism, among others. Second, the $\omega$-rule also enriches theories of arithmetic: PA supplemented with the $\omega$-rule can prove the consistency of PA and secureness completeness.

Hence, two intuitions on the semantic nature of the $\omega$-rule may arise. The first is that it in fact possesses arithmetical content, and this type of content now enters a proof system. The second concerns the infinitary nature of the rule: we might think that only the semantic side is fully equipped to deal with full infinity, where models may well be infinite mathematical structures. Thus, the infinitary aspect of semantics could be a property that can semantically pollute a proof system. Whether proof theory is indeed meant to only deal with finite proofs, is debated in the literature (this relates to the usual assumption that syntactic formulas need to be finite, which is also not a given — see for instance a discussion

in (Shapiro, 1991, §9.1.3)). Both suggestions provide avenues that can be explored more deeply as possible ways in which inference rules can be semantically polluted.

For now, note that infinity by itself is not a semantic culprit relative to the measure BR. It is easy to come up with trivially infinite rules (such as a rule that concludes $A$ from infinitely many copies of the same formula $A$), that do not violate invariance results under model equivalences. Still, we do note that such infinite rules do not have a (valid) formula interpretation, as the translation (if we assume compositionality) would need to result in an infinite formula. This particular guarantee for syntactic purity is then not satisfied in the usual settings.[3] The question remains in which exact sense the $\omega$-rule might satisfy BR. The $\omega$-rule ensures every truth in the standard model of PA becomes provable; implying that differing properties in non-standard models violate this rule. Yet relative to the 'baseline reference' of PA to, say, the standard model and all models isomorphic to it, the $\omega$-rule will simply hold (and allow better description of these models). Relative to a non-standard model, the $\omega$-rule will be false if it concerns a property that only the standard model satisfies. Hence, we expect that like Necessitation, relative to a big enough class of models, the $\omega$-rule may induce semantic pollution.

In short, the previous examples illustrate some preliminary insights into semantic pollution for inference rules. We see that a rule may satisfy BR on the level of validity (relative to a large enough class of models); and that infinity does not guarantee BR-satisfaction. In general, then, it is useful to see that translatability as a measure of syntactic purity and satisfying BR are not exact counterparts. Formula interpretations merely guarantee syntactic purity, while BR guarantees semantic pollution but does not capture *all* untranslatable formulas. Moreover, in future studies we might want to take a conceptually different approach towards analyzing semantic pollution for inference rules or for entire proofs. It seems that the type of semantic pollution that they would possess (if any), would be more syntax-independent (and so of a different nature) than the pollution of a proof-theoretic grammar. Recall that main examples of grammatical pollution relied on a clear separation from the local, valuation-dependent truth conditions of the modal language. This 'view' on Kripke models is something that we can less easily attribute to inference rules, which are based more naturally in validity (and so are in principle not local, nor valuation-dependent). We look forward to see possibly different conceptions of semantic pollution shape this ideal of proof on the levels of inference rules and formal proofs.

[3] Similar considerations can be given for cyclic or non–well-founded proof systems, that use infinitely many rule applications.

### 7.2.2 Type of proof system

For ontological purity, our measures were biased to natural deduction proof systems. Full purity does not require any adaptations when we switch to Hilbert-style or sequent-style proof systems; but secondary purity does, especially in the requirement for the 'threshold for pure formulas' (of Definition 4.3.2) occurring at some point in each proof branch. That is, we still want to be sure that at some point the proof restricts itself to formulas describing surrogates only. While we will not treat the cases of Hilbert-style and sequent-style calculi in full detail, we can illustrate preliminary workings of the derivation criterion in these settings by taking again Example 4.3.3 concerning a single-inference (by assumption fully pure) Q-proof, and a secondarily pure proof in $\mathsf{C}^2_{\mathsf{FO}}$. The fully pure inference was:

$$\dfrac{\forall x(x=x)}{0=0}\,\forall\mathrm{E}$$

The secondarily pure proof was as follows.

$$\dfrac{\dfrac{\begin{array}{c}\mathsf{C}^2_{\mathsf{FO}}\\ \vdots\\ \forall\mathbf{x}((\mathbf{T}(\mathbf{a},\mathbf{x})\vee\mathbf{x}=\mathbf{b})\rightarrow\mathbf{x}=\mathbf{x})\end{array}}{(T(a,b)\vee b=b)\rightarrow b=b}\,\forall\mathrm{E} \qquad \dfrac{\dfrac{\forall x(x=x)}{b=b}\,\forall\mathrm{E}}{\mathbf{T}(\mathbf{a},\mathbf{b})\vee\mathbf{b}=\mathbf{b}}\,\vee\mathrm{I}}{b=b}\,\rightarrow\mathrm{E}$$

Recall that the first derivation of $b=b$ in the rightmost branch does not satisfy the derivation criterion, and only the bottom instance shows that it can be done in a pure way.[4]

Now consider a Hilbert-style system, one where we only have the inference rule Modus Ponens, and the usual first-order axioms.[5] The secondarily pure proof can be seen to go as follows. Let $\delta(x) := T(a,x)\vee x=b$, and let all inference rules be instances of MP. Let the starred instances of the domain formula coincide to paste the two derivations together as one.

$$\dfrac{\dfrac{\dfrac{\forall x(x=x)\rightarrow b=b \qquad \forall x(x=x)}{b=b} \qquad b=b\rightarrow\delta(b)}{\delta(\mathbf{b})(*)}}{\dfrac{\dfrac{\begin{array}{c}\mathsf{C}^2_{\mathsf{FO}}\\ \vdots\\ \forall\mathbf{x}(\delta(\mathbf{x})\rightarrow\mathbf{x}=\mathbf{x})\end{array} \qquad \forall\mathbf{x}(\delta(\mathbf{x})\rightarrow\mathbf{x}=\mathbf{x})\rightarrow(\delta(\mathbf{b})\rightarrow\mathbf{b}=\mathbf{b})}{\delta(b)\rightarrow b=b} \qquad \delta(\mathbf{b})(*)}{b=b}$$

[4] Although we do not do so here, this suggests that it might be interesting to further analyze the way that the derivation criterion encourages (perhaps even requires) cyclicity.

[5] In the first-order setting, quantifier rules can be common for Hilbert-style systems. However, the setting where the axiom-rule ratio includes the largest number of axioms and the least number of rules may show the differences with respect to natural deduction most clearly.

We see that the definition of 'pure formulas' (as in Definition 4.3.2) needs to be enlarged in this setting, in order to include some 'good formulas' (as in Definition 4.3.1) as well. Take for instance $\forall x(\delta(x) \to x = x) \to (\delta(b) \to b = b)$, which is a 'good' formula (and not a 'pure' one), but is also an axiom of the proof (in the natural deduction setting, only assumptions that can later be discarded are examples of leaves that are 'good'). The Hilbert-style focus on axioms makes it harder to separate 'pure' formulas (such as $\forall x(\delta(x) \to x = x)$) from their 'good' recombinations. Additionally, and related, the loosened threshold of purity occurring in a proof generally also shifts upwards towards the leaves of the proof, as the Hilbert-style axiom schemas allow instantiations by (recombinations of) $\delta$-relativized sentences. On the contrary, in natural deduction proofs, the Hilbert-style logical axioms are built into rules, so that the pure formulas always really occur in the middle of a formal proof, or as an assumption. In short, and seeing that the rule MP itself does not need a restriction, derivation criterion for secondary purity is really generalized to act on the grammar.

Consider now a (shared-context) sequent-style system for first-order logic (after (Negri and Von Plato, 1998), we let all instances $\Rightarrow a = a$ be axioms). The secondarily pure proof can be depicted as follows.

$$\dfrac{\begin{array}{c}\mathsf{C}^2_{\mathsf{FO}}\\ \vdots\\ \Rightarrow \forall \mathbf{x}((\mathbf{T}(\mathbf{a},\mathbf{x}) \vee \mathbf{x} = \mathbf{x}) \to \mathbf{x} = \mathbf{x})\end{array} \qquad \dfrac{\dfrac{\dfrac{\Rightarrow b = b}{\Rightarrow T(a,b) \vee b = b}\vee\mathrm{R} \qquad b = b \Rightarrow b = b}{(T(a,b) \vee b = b) \to b = b \Rightarrow b = b}\to\mathrm{L}}{\forall \mathbf{x}((\mathbf{T}(\mathbf{a},\mathbf{x}) \vee \mathbf{x} = \mathbf{x}) \to \mathbf{x} = \mathbf{x}) \Rightarrow \mathbf{b} = \mathbf{b}}\forall\mathrm{L}}{\Rightarrow b = b}\,\mathrm{Cut}$$

Several observations can be made with respect to the derivation criterion in this setting. First, recall that in sequent calculi with mathematical axioms, the Cut rule is commonly not eliminable, making only free-cut elimination feasible. The application of Cut in this proof shows how the Cut rule connects two separate parts of the proof. The left part starts from powerful axioms, and works down towards a relativized (pure) axiom, marking the onset of a pure derivation. For the right part, we might choose whether to take a top-down, or a bottom-up perspective on the rules. The top-down perspective seems to suggest that the purity guarantee on the right side also only kicks in at the premise of the Cut rule, where (thinking of the sequent arrow as representing derivability) the sequent $\forall x((T(a,x) \vee x = x) \to x = x) \Rightarrow b = b$ represents a pure derivation of $b = b$. As for the part of the proof above the Cut premise, we might think that we do not have enough information about the 'derivability context' of the relevant formulas yet. The atomic formulas at the leaves are gradually built into more complex formulas, either restricted to referring to surrogates, or not. After building up some complexity, they might form instances of a pure formula, or instead something more powerful. This conceptual perspective makes the derivation criterion less interesting, as it secures only the purity of the last inference, that of the Cut rule.

More natural might then be to take on a bottom-up perspective of the rules in the right part of the derivation. While the left part of the derivation functions to derive a pure (relativized) axiom, the right part then serves to display how this relativized axiom is *used*, i.e. how we can reduce its complexity in a pure way. On this perspective, the entire right part may be considered to find itself among good formulas. When we read the tree bottom-up, the rule $\forall L$ (bottom-up) can be seen to perform quantifier elimination to the left side of the sequent, which is a safe operation purity-wise. More decomposition brings us to the building blocks $b = b$, which still originates from the pure cut formula. That is, decomposing a pure derivation does not give us anything extraneous. Hence, this understanding says that a sequent calculus proof may efficiently outlaw any extraneous formulas to a particular part of the derivation that works from mathematical axioms; the other part of the derivation then takes a derived formula, and brings it back to simpler, still pure, atomic formulas. In our example, the premises of the Cut rule may then be considered to mark the pure threshold, and to induce further upwards purity of the proof on the right side.

Additionally, since the bottom-up reading of the proof only reduces complexity of formulas (aside from possible uses of the Cut rule), no additional restrictions on inference rules need to be placed, once we work from a pure formula upwards. The cut formula is thus the most important source of possible impurity, and needs to be restricted by requirements of 'goodness' or 'pureness'. Recall that in (Arana, 2009) (see also the remarks of Section 3.2), it was investigated whether cut elimination could 'purify' a proof, assuming that the cut formulas themselves can contain extraneous content, precisely because they violate the subformula property. The idea was rejected there, mainly because of the need for free-cut elimination for first-order mathematical theories, reducing the subformula property to one where formulas may also be subformulas of the (mathematical) axioms. It is a nice additional insight that we now see that the subformula property relative to a *pure cut formula* (as we defined in Definition 4.3.2) in fact is sufficient for purity. Our example suggests that the cut formula may be seen as the pure threshold, derived from stronger mathematical axioms — 'after' this, a pure derivation is guaranteed, as shown by the (cut-free) part of the derivation that works its way back to the logical axioms.

We repeat that for a thorough understanding of the derivation criterion for Hilbert-style and sequent-style calculi, the precise formal details should be worked out elsewhere. Finally, regarding the type of proof systems used for defining semantic pollution, we restricted ourselves to sequent calculi and generalizations. The same grammar requirements can be imposed on natural deduction and Hilbert-style proof calculi, easily. However, they seem especially suitable for sequent calculi, as the generalizations of the data structures used there are most sensitive to grammatical pollution. In natural deduction and Hilbert-style calculi, extended proof-theoretic languages are less common, although labeled formulas may cer-

tainly also occur there (see e.g. (Basin et al., 1998)). Then again, it might turn out that natural deduction and Hilbert-style systems are more prone to other types of semantic pollution, for instance occurring on the level of inference rules, or entire derivations. A final interesting generalization for semantic pollution would be to consider the setting of semantic tableaux (see, e.g. (Priest, 2001)), where proofs work from the axioms but also the negation of a conclusion, towards a contradiction. Semantic tableaux seem have a more specific connection to the model theory, by providing a method to identify a specific (counter)model from a formal proof. This may indicate a different type of semantic pollution, appropriate for this style of proof system.

### 7.2.3 Background logic and logical semantics

Changes in the background logic of our case studies also provide new contexts in which to make the formal measures precise. In particular, we defined ontological purity with respect to classical, first-order logic. A change to intuitionistic logic would raise some conceptual as well as technical questions. Conceptually, intuitionistic theories adapt the style of reasoning and proving done on an ontology, but perhaps preserve most aspects of the ontology itself.[6] Technically, especially regarding the derivation criterion, more open questions arise. In the intuitionistic world, there less Visser (1997)-style interpretations are expected to exist, because commutativity with of the interpretation translation with intuitionistic connectives works less well. This suggests that a mathematical structure will be occupied by less intuitionistic mathematical domains than classical ones, resulting in a slightly stricter notion of secondary purity for formal proofs. This does not have to be a bad thing: the fewer interpretations that work may instead be more natural, as the very technical and complex interpretations are now ruled out. Note that these changes only really make a difference for purity results of formal derivations, however. For informal proofs, there are built-in requirements on naturalness already, so that secondary purity for informal proofs does not change just because there are *less* interpretations into intuitionistic versions of theories — only if interpretations between such theories stop existing altogether, the informal purity result would also stop holding.

Finally, different notions of reducing theories to each other (such as realizability) exist in the intuitionistic setting, compared to the classical situation. It would be interesting to see if any such notion could lead to a different notion of surrogacy, or to different derivation criteria (the same holds for second-order versions of mathematical theories and notions of equivalence).

As for semantic pollution, we considered classical (propositional) modal logic in our case study. Changes to intuitionistic, or higher-order logics lead to different

[6] With the exception of more extreme constructive viewpoints: for instance, *strict finitism* would arguably strongly reduce the complexity of ontologies, in particular only allowing finite ones.

truth conditions of logical constants, and so to a shifted baseline reference relation, including new (variants of) model equivalences. A change of background logic may naturally come with a change of logical semantics, as well (although the same logic may already have various types of semantics itself). All these new baselines provide suitable case studies for semantic pollution with respect to various proof-theoretic languages. As mentioned in the previous chapter, likely candidates for semantic pollution may be 'neighborhood proof systems' (Dalmonte et al., 2018; Negri, 2016) (introducing besides 'world labels', for instance also 'neighborhood labels' or even set-theoretic predicates), as well as a proof system for intuitionistic predicate logic (Baaz and Iemhoff, 2008) (introducing axioms that include symbols for among others a set of Kripke worlds and the preorder on them, as well as the first-order domains attached to Kripke worlds).

Taking the baseline reference relation of our case study as fixed for the moment, other interesting questions arise. Suppose for instance, given this baseline, that we change to neighborhood semantics. Can the level of pollution of the same proof-theoretic syntax now change? The notion of formula interpretation is independent of semantics (as we defined it based on provability and logical entailment only), and so syntactically pure formulas will at least stay syntactically pure under such a change. That is, they retain an outlook on the new type of semantics similar to that of the object language. But perhaps syntactic elements without a formula interpretation can now lose or gain (levels of) semantic pollution — this might happen when a formula obtains a new truth condition with respect to the different semantics, which satisfies different properties than before. We look forward to more insights into these topics. For instance, relative to neighborhood semantics, $x : A$ will arguably retain its truth condition $x \vDash A$, and is likely to still violate invariance under neighborhood model equivalences, which will not immediately take into account a label assignment function. However, it seems already more of a challenge to provide a truth condition for $xRy$ relative to neighborhood semantics, and analyze its level of semantic pollution there.

### 7.2.4 Conclusions

We see various possibilities of adapting the measures for both ontological purity and semantic pollution to other proof-theoretic and logical contexts. The interesting generalizations for ontological purity seem to mainly concern other types of proof systems, while semantic pollution seems also worth generalizing to different ingredients of a proof system, and to different types of logical semantics. This difference arises from the fact that the measures for purity pose an order restriction on formal proofs, which manifests itself differently for various types of proof systems. On the other hand, our focus for semantic pollution was restricted to proof-theoretic grammar, leaving other characterizations open. We hope to obtain more future insights into the relation between the formal measures for ontological purity and semantic pollution, as well as their generalizations.

## 7.3 Reflections on formalization

In this final section, we address the insights we gain into the notion of formalization. In Chapter 1, we introduced the possible advantages and disadvantages of formalizations of a pre-theoretical concept. Here we want to see more specifically, for the three different types of formalization we carried out (see Section 1.3), what they contribute to.

Recall from Chapter 1 the various theoretical virtues and dangers of formalizations outlined by Hansson (2000). We describe the main patterns of virtues and dangers found for formalizations of proof ideals for informal proofs (on purity and explanation), the formalization of purity for formal proofs, and the formalization of semantic pollution for formal proofs. We end by shortly reflecting some more on what we have learned about the formalization of an ideal of proof in general.

### 7.3.1 Ontological purity and explanation for informal proofs

In Chapter 2 we already analyzed some aspects of the type of sharpening done by models of ideals for informal proofs — here, we emphasize some further aspects of these formalizations. First, models of explanation and purity for informal proofs rely on *intuitions* in an essential way, which characterizes their virtues and dangers. They certainly serve to sharpen or elucidate a pre-theoretic conception of an ideal. However, the level of sharpening is itself also of an informal nature. Sharpening is achieved by introducing new terminology that has more conceptual content than the existing pre-theory, but that still remains indeterminate with respect to various aspects; hence a reliance on intuition persists. For instance, models of purity sharpen the notion of 'content' of a theorem, and how to restrict a proof to it by introducing new terminology (*topic*, *ontology*, and so on); models of explanation sharpen the notion of an explanatory relation between a proof and a theorem (by introducing terms like *argument pattern*, *characterizing property*, and so on). This ensures that sharpenings can manifest themselves in widely different ways.

Dependence on intuitions has advantages, as it avoids possible oversimplification by reducing everything to one formal ingredient; and prevents the formalization from only capturing very specific examples of proofs satisfying the ideal. Simultaneously, the ingredients that *are* fully sharpened manage to provide new insights into the ideal. For instance, we specified a certain variant of purity, where we suggest that a mathematical structure can be referred to by a formal set of axioms, its definitional extensions, and their interpretations into other theories. This surely narrows down the pre-theoretic notion of 'content of a theorem', and shows a way in which we may conceive of purity as having degrees. Additionally, such a sharpening provides new emphasis on philosophical problems such as: when can two formulations of theorems be considered the same? What does it mean

more specifically to be a (natural) formalization? More specifically: what is the boundary between a property being formally definable by a theory, and for this definition to be a natural one?

The reliance on intuitions also relates to the dangers of the formalization. The type of 'formal oversimplification' that is avoided by leaving room for intuitions, is regained when considering a more conceptual type of oversimplification. That is, by defining an informal, relatively open type of concept, we require our intuitions to always have a say in how to fill it in. But consider the crystallization of content into an ontology: one simply cannot always have clear intuitions on an ontology for a given theorem (for instance, recall examples on meta-theoretic theorems like the consistency of theories; or set-theoretic theorems that could have wildly different (sizes of) ontologies associated to them). In those cases, what its content is supposed to be is unclear. Similarly, the step from an ontology to a formal theory can be oversimplified, as mentioned in Chapter 3. That is, a formal theory will never perfectly capture an ontology, for instance because its axioms in the formal setting turn out more or less powerful than previously thought (consider Gödel's Incompleteness Theorems); or, because the set of formal primitives do not completely capture the intuitively important objects and operations. This danger is also found in models of other variants of purity and explanation for informal proofs (for instance, in selecting the set of operations mentioned by a theorem, an argument pattern, a characterizing property, and so on). While some key examples may fit these conceptual terms well, in other cases they may be highly oversimplified. Although we advocate these sharpenings as merely specifying a variant of an ideal within a pluralist setting of various formalizations, the danger remains that they are taken too universally, and that we let them 'blunt' the philosophical debate by ignoring the more refined intuitions underlying ideals like purity.

Furthermore, we should be aware of the purpose of such formalizations. Because the ingredients of models of ideals for informal proofs are not calibrated for application to proofs where no clear intuitions exist regarding the ideal (or ingredients of the model), these models seem to primarily serve providing more theoretical insight into the nature of an ideal (rather than aiding the practical verification of whether or not proofs possess an ideal). Given certain intuitive examples (not) possessing the ideal, such models provide more theoretical insight into what underlies these intuitions. Applying the model to 'neutral' examples where no intuitions about the ideal exist, may still provide more conceptual insight into the model itself. It shows us which aspects of a model are well-defined, and which ones need a tolerant interpretation and rely on intuition. And it provides insight into which aspects sharpened by the model are present and absent in the example, providing a sense of how it relates to and departs from some model of the ideal. Our study in Chapter 2 for instance helped show how the sharpenings of two different proof ideals interact theoretically, showing whether our intuitions for these ideals are in fact based on the same theoretical properties, or not. A

shared underpinning of multiple ideals would support the idea that these theoretical properties are important in clarifying what makes mathematics revealing (as ideals of proof guide mathematical practice). It would also suggest that such shared properties are key to successful formalizations. The inconsistent results of Chapter 2 concerning purity and explanation, however, suggests that stable properties underlying formalizations of proof ideals may be hard to find.

### 7.3.2 Ontological purity for formal proofs

Consider now more specifically the model of ontological purity for formal proofs, given its model for informal proofs. We might say that this model serves less to elucidate a pre-theoretic conception, and more to provide knowledge on the formalization of informal proofs into formal proofs. The question here is, given an informal proof that is ontologically pure, which formal proofs should be given a level of purity? We focus on some specific aspects relating to the formalization of ontological purity.

First, the main sharpening of ontological purity for formal proofs concerns its relation of the 'informal use' of an ontology in a proof to syntax restrictions for formal proofs (to any syntax of the context theory, and to a syntax restriction (in a particular order) in theories interpreting the context theory). This sharpening shows that there is a fine balance between which informal aspects of a proof can be preserved, and which are ignored in the formalization. For instance, while an informal proof is 'really' able to restrict its ontology to surrogates, by simply not mentioning any other objects or operations, our natural deduction proofs need to mention the unrestricted axioms of interpreting theories in order to provide a 'gapless' formal proof. Hence, a choice arises concerning how one formalizes use of surrogate ontology in a full formal proof — we chose to do this with an order constraint on syntax occurring in the formal proof (other ways may exist). Similarly, as long as syntax belongs to the language of the context theory or to the 'good' formulas of an interpreting theory, we do not take inference rules as affecting the ontology. This means that no matter how complex a syntactic string becomes through manipulation by inference rules, such that it conceptually does not correspond anymore to a natural concept, or such that it might conceptually be more similar to a concept from a different ontology — we will still consider it pure, as it will have an acceptable ontological interpretation.

We see some dangers of the formalization of ontological purity for formal proofs concerning oversimplification and sensitivity to technical artifacts, as before. As mentioned above, the requirements of pure formal derivations are perhaps loose (by accepting all syntax produced by a set of axioms), but also strict (by the specific use of the interpretation translation, and the specific requirements of the derivation criterion — both might be generalizable). This undoubtedly creates some distortion of the conception of ontological purity for informal proofs that we started out with. However, note that the formalization of ontological purity for

informal proofs was a sharpening already of the pre-theoretic conception of purity generally, so that its subsequent transferral to the setting of formal proofs can (as a sharpening of a sharpening) be considered relatively accurate. That is, relative to the pre-theoretic conception of purity, generally, the formalization of ontological purity for formal proofs provides various distortions and oversimplifications — but relative to the model for ontological purity for informal proofs, we are left with fewer gaps to fill. The gaps that are left, such as the choice of ontology, and notion of '(natural) formalization', are included in the formalization as intended remaining open texture. Then, once these choices are made, ontological purity for formal proofs is entirely sharpened.

Further, the formalization of an ideal of proof to the setting of formal proofs may come with supplementary formalizations of related concepts. For instance, secondary ontological purity for formal proofs provided an accompanying sharpening of the notion of a *proof simulation* (by Theorem 4.4.3). This can be seen as a proof simulating not only the ontology of the context theory by surrogates, but also by simulating certain proof steps (and perhaps the proof strategy) of a formal proof. While this is an aspect of formal proofs that is too fine-grained to be relevant for ontological purity, the concept might provide clarity in more formal settings. The formal tools used in obtaining these sharpenings additionally give rise to several open questions. For instance, how can we characterize the proofs satisfying a derivation criterion that are *not* a simulation? More generally, what other interesting ways of simulating proofs are there, and what proof aspects do they fix? What other translations between theories induce some (variant of a) derivation criterion? And what philosophical properties do (proof) translations really preserve, and which ones do they not preserve? These are just a few examples of many others, and we leave the search for precise answers to these questions to further research.

### 7.3.3 Semantic pollution for formal proofs

Finally, we turn to the model of semantic pollution for formal proofs, which serves to elucidate a pre-theoretic ideal that already concerns formal proofs. Because the notion of a formal proof is well-defined, the pre-theoretic ideal may itself be considered a bit more precise already. If we consider the informal description of semantic pollution as 'a semantics invading a proof system', we are already clear about the nature of the terms 'semantics' and 'proof system'. It is the notion of 'invading' that is in particular need for formalization.

Like ontological purity, syntactic purity and 'extraneous' semantic elements are made precise by differences in language strength and formal notions of sameness. Our sharpening says that semantic pollution is about a certain simplicity of syntax, and that stronger syntax that violates invariance results under model equivalences can express 'too much' about some (type of) semantics. Again, the formalization leads to some interesting further questions, including: how can we measure the

power or effect of an individual operator, when it is added to a language? How can we better characterize the gap between being untranslatable to a language, and violating invariance results under model equivalences? Furthermore, how do syntactic purity and semantic pollution relate to informal proofs? We might envision specific cases where a similar informal variant of these properties hold: an informal proof in the field of proof theory may be semantic (using the model theory), or syntactic (using just proof-theoretic tools). A model-theoretic proof of a syntactic property such as cut elimination might be considered 'semantically polluted' in this case. However, the informal counterpart of the ideal seems to have a slightly different conceptual flavor than semantic pollution for formal proofs, and for informal proofs, it might in fact relate more generally to ontological purity. This is something that deserves further exploration.

A specific aspect of the sharpening of the base requirement, concerns our understanding of 'simple' syntax. The formalization shows that language strength is not all-determining for semantic pollution; an important factor is also the level of *connection* to a language. To illustrate this, consider the following example.

**7.3.1 Example.** Consider first-order modal logic as the baseline syntax, relative to first-order Kripke semantics and a labeled proof-theoretic language. Now take a labeled formula, such as $x : \forall a(\Box a)$ or $xRy$. A (constant-domain) first-order Kripke model is commonly a tuple $(W, R, D, I)$ where $W$ is a set of worlds, $R$ is the accessibility relation on worlds, $D$ is the constant domain associated with each world, and $I$ is a function assigning to each pair of an $n$-ary predicate and a world, an $n$-tuple of elements of $D$. Now, even though the object language is of first-order strength, the model is not yet equipped to provide truth conditions to labeled formulas, because $I$ does not interpret labels. Hence, the truth condition of $x : \forall a(\Box a)$ and $xRy$ will still need a model $M = (W, R, D, I)$ to be augmented by a label assignment function $\tau : \mathsf{Lab} \to W$. Furthermore, if we consider any usual model equivalence for $M$, it clearly does not include any conditions on $\tau$. Hence, as before, new model equivalences will have to be defined for models $(M, \tau)$. For weak model equivalences that just 'add $\tau$ on top', it is easy to see that labeled formulas will still satisfy BR. Semantic pollution thus still occurs.

Hence, in a sense, semantic pollution is, besides the strength of a proof-theoretic language, also about the connection between the model-theoretic interpretation of elements of this language. The formula $xRy$ may be thought of as similar in strength to the first-order object language, but semantic pollution can still occur because the interpretation of the set of labels is separated from the set of first-order variables. Recall that we also saw an example of the converse in the previous section on generalizations: the situation where the object language was not strong enough to translate a proof-theoretic element (take a trivial infinite rule with infinitely many copies of the same premise), but where this did not result in semantic pollution. There, all the ingredients of the (infinite) representation of the

rule, were still compatible with the object-language perspective on the model theory. These examples further clarify what the formalization of semantic pollution focuses on, and what it does not.

Finally, the formalization may be seen to suffer from dangers of similar nature as before. Oversimplification of intuitions is one: based on different interpretations of the open texture of the pre-theoretic ideal of syntactic purity, one might prefer different a 'calibration' of the formalization of semantic pollution. For instance, one might want semantic pollution to be more sensitive towards untranslatability, less sensitive to the 'connectedness' between syntactic elements, or even more sensitive to the use of syntax, e.g. their manipulations in formal proofs. Or, within our focus on the grammar of a proof system, semantic pollution might instead be defined in terms of a baseline reference relation that is more tolerant than just the perspective of the object language. The definition of such formal variants of semantic pollution would be nicely comparable to our original variant, and it might lead to a family of formal notions of semantic pollution.

### 7.3.4 General remarks on formalizations of proof ideals

A few final remarks concern some more general properties of a formalization of an ideal of proof (which require a more detailed analysis for definite conclusions). A main observation concerns the idea that pre-theoretic properties can be formalized with varying levels of 'fixedness', or levels of *remaining open texture*. By this we mean that the formalization can contain different levels of built-in flexibility for an agent, relative to how the open texture of certain aspects is accommodated. For instance, for the formalization of the notion of computability by Church's Thesis (see Chapter 1), essentially nothing is left for an agent to interpret for themselves. Indeed, we might think that a formalization is supposed to turn all aspects of the pre-theory into formal counterparts. But our case studies have shown that sometimes we might prefer only formalizing *some* properties of the pre-theory, or only *partly* formalizing the properties of the pre-theory. In order to be more aware of this possibility, we might want to categorize formalizations according to their level of 'remaining' open texture. After all, a formalization is only satisfactory if it fixes as well as leaves open the right notions.

At first sight, (at least) four levels might be distinguished. Rigid formalizations of pre-theoretic ingredients may be called *fixed*: any existing pre-theoretic open texture is completely taken away. That is, an agent $\alpha$ using such a formalized model has no choice whatsoever in determining the ingredient of the model, as the model determines what the formalization becomes. In our formalization of ontological purity, the notions of interpretations and definitional extensions have this fixed character (relative to a context theory). Similarly, the property of violating invariance results under model equivalences is fixed (relative to a reference relation), as well as our chosen properties of globalness and valuation independence. These properties make up automatic components in the evaluation of (ontological

or syntactic) purity of formal proofs.

Ingredients that are slightly less rigidly formalized we might call *semi-fixed*. For instance, this could include the case where a pre-theoretic notion is formalized into a completely determinate type of formal counterpart — i.e., $\alpha$ cannot choose the nature of the formalized object (or property).[7] However, $\alpha$ has freedom in choosing a precise instance of this object or property, that she thinks fits the formalization at hand best. In our formalizations, this may involve the particular choice of the baseline reference relation (including a choice of context theory, and for semantic pollution, a choice of syntax and semantics). Additionally, the notion of an 'operational domain' (from Kahle and Pulcini (2017)'s operational purity) may be seen to satisfy this level, since the nature of a numerical domain is clear, but there is a degree of freedom in deciding upon one.

Other levels provide less ground to stand on for formalizations. A, say, *semi-flexible* formalization can be given of a pre-theoretic ingredient, when there exists a relatively stable, shared intuition on its nature — yet when applying the formalization in practice, the precise details need to be gauged per individual example. That is, although the ingredient is characterized by a strong basic understanding (among the mathematical community), the variety of practical examples requires an agent $\alpha$ to form a unique perspective on each individual formalization. Several ideals of proof can be considered to have ingredients of this nature. The notions of (surrogate) ontology, and structure, are arguably examples of such notions for ontological purity. The notion of operation (for operational purity), dependence (in Steiner explanation), and abstraction (in Pincock explanation) seem also suitable examples of semi-flexible formalizations. Surely there is a wide, shared understanding of what these notions are supposed to mean. However, reducing them to a fixed counterpart, even a single type of object or property, seems highly unlikely to do justice to the variety of ways in which proof ingredients can satisfy them.[8]

Finally, a formalization may include ingredients that are such 'loose' sharpenings, that we could call them *flexible*. Here, a suggestive term might be given to the formalized ingredient, but it lacks the stable, shared understanding that was present in the previous level. Our agent $\alpha$ has the freedom to choose precisely what type of formal object this becomes, as well as which precise instance of this object she chooses: as long as she can argue in some convincing way that these choices satisfy the suggestive requirements of the formalization. In topical purity, while the notion of 'topic' and 'co-finality' might be evaluated as semi-flexible, there is arguably less shared understanding surrounding these notions. That is, the type of things that a topic can be are wide-ranging, and co-finality as a property is quite sensitive towards different interpretations. Other properties perhaps falling

[7] Such as a set, a function, a formal theory, a language or a model, any type of thing that we deem rigorously defined.

[8] Note that, although we for instance relate an ontology to a formal theory, we do not suggest that they are the same types of things. An ontology is considered to be a separate entity from a formal theory, the latter of which is considered to *refer* to the former.

into this category include Steiner's 'characterizing property', as well as Kitcher's 'argument pattern'.

Hence, a formalization can be seen as coming with an optimal 'level of specificity', or balance between intuition and rigidity, that allows it to work best in practice. Overspecification relates to many dangers of formalization, and can quickly be seen to make formalizations useless, and separated from pre-theoretic intuitions. For our examples, we see the trend that formalizations for informal proofs leave some level of open texture for each ideal. Even for semantic pollution and syntactic purity, as properties of formal proofs (concerning the highest level of determinacy), the initial choice of the baseline reference relation leaves some open texture. For instance, labels are not 'polluted by definition'; they are polluted relative to the baseline chosen. Similarly, a proof of a theorem is not pure or impure in some absolute sense, but only relative to a choice of ontology and context theory.

There should be many other interesting general properties of formalizations for ideals of proof that we have not considered closely. For instance, formalizations may be differently suitable for defining levels or degrees of the property that they formalize. In general, we expect that a formalization with less remaining open texture provides more opportunities for defining levels. Namely, if all ingredients of the formalization are rigid, it is easier to play with the chosen variable settings by tuning them slightly differently, and to see what properties of the pretheory remains. Furthermore, a general property of formalizations of proof ideals is whether formalizations tell us anything about the relation between informal and formal proofs. The relation between informal and formal proofs can be seen as inherently connected to the problem of *proof identity*: the former might be interpreted as asking which informal proofs are 'the same' as which formal proofs. In Chapter 1 we mentioned that one way to distinguish proofs conceptually is by seeing which ideals of proof they (do not) possess. This implies that formalizations of ideals of proof might provide more insight into (conceptual) proof distinction, and so into the relation between informal and formal proofs.

In general, formalizations at the level of informal proofs (as in Chapter 2) seem to give at least a boost in specifying sharpened informal properties that are relevant for conceptual proof distinction — they can be used as guidelines for bringing an ideal to the formal setting. Formalized ideals of formal proofs (such as semantic pollution) more easily provide distinctions between formal proofs, but these properties are not guaranteed to correspond to anything meaningful on the informal side. The only ideal defined for both informal as well as formal proofs was ontological purity, providing a first way to link the two. We look forward to gaining more insights into the exact connections between the two types of proofs. For now, in line with pluralism on formalizations, we suggest that a context- or property-relative attitude to proof identity, and to the relation of informal to formal proofs, is useful — implying that criteria of identity are variable relative to properties of interest.

## 7.4 Conclusion

In this chapter, we have considered the formalization of ideals of proof from a more general point of view. We considered the likenesses and differences between the formalizations of ontological purity (for formal proofs) and semantic pollution, and we saw several potential generalizations of their criteria. Finally, we discussed the value and nature of the specific formalizations. Plenty of directions for future research can be seen to emerge here. While proof systems are ever-changing, we look forward to new ones coming into existence, and to formalizations of ideals of proof being developed in new settings.

# Conclusion

In this dissertation, we have explored three ideals of proof, and attempted to compare them more closely and define several of their formalizations. In particular, Chapters 3 and 4 introduced a new variant of purity of mathematical proofs into the literature, that of ontological purity, and examined under which criteria formal natural deduction proofs can be said to satisfy a full and secondary level of this type of purity. Secondary purity was argued to deserve its name by securing a reference relation to surrogate ontological content through criteria relying on the interpretation translation, thereby preserving the structural content of a theorem. We thus provide a first characterization of purity for formal proofs, while showing that the particular philosophical viewpoint on mathematical content that is adopted, can affect the resulting purity outcomes.

Ontological purity also occurred in the broad comparison of models for purity and explanation of informal proofs in Chapter 2. There, we investigated the behavior of each model when applied to a standard proof of the Infinitude of Primes (displaying features of purity), and a standard proof of Pythagoras's Theorem (displaying properties of explanation). Such a case study showed that the intuitions underlying a single ideal of proof are multi-faceted and can lead to widely different sharpenings. While the geometric notion of similarity occupied an explanatory role in the proof of Pythagoras's Theorem, it was able to introduce epistemic, operational as well as (in a geometric setting) ontological types of impurity according to various models. The 'pure' arithmetical tools used to prove the Infinitude of Primes turned out to be compatible with Kitcher's unifying argument patterns, but showed no clear characterizing property in the sense of Steiner's model of explanation, nor an abstraction difference as intended by Pincock's model. Hence, the comparative study of purity and explanation reinforces the idea that, although these ideals may share a conception where pure and explanatory elements possess a certain basic simplicity, they differ in richer sharpenings that allow purity and explanation to stem from more wide-ranging conceptual properties. The study also featured the more tolerant workings of our notion of ontological purity, that allowed purity of the proof of Pythagoras's Theorem as long as the theorem is conceived of as referring to an ontology including arithmetic.

We made the switch to a discussion of proof systems and Kripke semantics for modal logic in Chapter 5, in order to ready ourselves for the formalization of semantic pollution in Chapter 6. The preparation consisted of an overview of

the display language, the labeled language and the hybrid language, as well as a definition of various types of Kripke model equivalences that incorporate requirements on the label assignment function. After providing an introduction to the notion of semantic pollution and possible formalizations, we then introduced a first elaborate conceptual understanding of semantic pollution in Chapter 6. In particular, we suggested that a semantics can be said to 'invade' a proof-theoretic language when formula types from this language make more fine-grained distinctions between models than the object language does. Hence, the base requirement for semantic pollution was formalized as formula types from proof-theoretic languages violating invariance results under model equivalences. Additionally, taking the local and valuation-dependent properties of the modal language as characteristic of the modal perspective on Kripke models, we suggested that globalness and valuation independence, on top of the base requirement, indicate even stronger levels of semantic pollution. This led to a first systematic classification of the levels of semantic pollution of various proof-theoretic languages, in which labeled languages displayed a high level of semantic pollution.

These case studies can also be seen to contribute to the broader question of what is a 'good' mathematical proof, and a 'good' design of proof systems. This is a topic Chapter 1 showed comes with many motivations, ranging from desirable epistemic qualities of a mathematical proof, to technical motivations from proof theory itself, but also relating to proof-theoretic semantics, categoricity, and so on. Whatever ones goal for a proof or a proof system, we advocate generally a conscious design of proof systems relative to that goal. By this we mean that there should be an awareness of the plurality of motivations in the field, and a recognition that the design choices made for a specific goal, may affect other goals (even if one may not have these other goals oneself).

Finally, Chapter 7 provided a preliminary more general perspective on our case studies. We observed that the formalizations of ontological purity and semantic pollution are alike in taking a reference relation as a starting point, which forms the context relative to which they can be measured more formally. On the other hand, we emphasized that our interpretation of these pre-theoretical ideals were based in the referents for ontological purity, but based in the reference relation itself for syntactic purity. We also saw that the ideals we considered benefit from keeping a level of open texture within the final formalization. Finally, we believe that the formalizations of ontological purity and syntactic purity are suitable for generalizations relative to various aspects, including the ingredients of a proof system, different types of proof systems, and differences in background logics and logical semantics. All these topics lead to various interesting possibilities for future research.

All in all, we encourage a pluralist and open-minded attitude towards the formalization of proof ideals, where ones preferred pre-theory offers guidance on an appropriate level of remaining open texture in a formalization, and where formal tools provide clarity where possible.

More broadly, this work can be considered to touch on various higher-level topics. We end our contemplations, and this dissertation, by highlighting a few of them below.

**'Natural' syntactic codings and translations.** Although we provided case studies of formalizing ideals of proof, formalization in mathematics more generally remains an obscure, quite impenetrable notion, where lots of insights may still be gained. The type of formalization of (mathematical) notions into formal theories or formal proofs involves, ultimately, a choice of syntactic representation of this content. As seen in the case study of ontological purity, the distinction between 'formalizable' and 'naturally formalizable' is for instance certainly one that deserves more attention, and which may lead to a better understanding of formalization generally. In this dissertation, we took this distinction as a guideline (evaluated by the standards of the mathematical community) for when an informal proof is 'fully' pure, and when it is 'secondarily' pure — more clarity on natural versus more 'artificial' formalizations would clarify the workings of ontological purity on the level of informal proofs. On the side of formal proofs, formalization had already occurred, and the interpretation translation was allowed to induce secondary purity in any (valid) way. Still, even there, more insight into the distinction between 'natural' and 'artificial' interpretations between theories might make the notion of surrogate content more tangible, and would reduce the distance between informal proofs using surrogate content, and their formalizations relative to an interpretation translation.

More specifically, the naturalness of the *primitives* as occurring in a set of context theory axioms has been key for ontological purity. Definitions of mathematical concepts that are not directly given by the primitives may however quickly involve layers of 'codings' where basic syntactic elements function as representations of various different concepts (PA's syntax for natural numbers can code, for instance, complex arithmetical properties, but also concepts such as negative numbers, set-theoretic membership, and so on). Clearly, however, syntactic complexity cannot function as a measure of a conceptual notion of artificialness: languages with a small signature (such as $\{\in\}$) simply need to use high syntactic complexity (and more codings), and even for simple concepts, than languages with a rich signature. Recall also for instance Example 3.4.4, that commented on the formalization of an uncountable set in PA, and whether representing just one element of this set can be considered a 'formalization' of the entire set. These issues raise the question: how distanced may representations of mathematical content be from their original concept in order to still be a 'formalization'? And how can we define a notion of syntactic 'naturalness' that is independent of a signature's richness, and determined not just by syntactic complexity?

**Informal and formal proofs.** Our case studies certainly touched on the question of how informal proofs relate to formal proofs, and in particular, how each may be said to approximate the other. For ontological purity, we suggest that a (possibly surrogate, and structural) ontology of a theorem is a property that can be described for both informal proofs, and for formal proofs. Hence, considering particular (philosophical) properties of proof as a fixed property under formalization, can be an effective strategy for linking informal proofs to formal proofs.

However, again, the question remains how informal and formal proofs relate to each other in a more general sense. We note that several approaches to answering this question exist in the (growing) literature on this topic — one of them is to work with more refined levels of formality. For instance, besides the formal proofs in proof systems as we know them, and informal proofs as certain natural language descriptions of proofs, "it is also possible to imagine higher-level proofs that are nonetheless presented in a language that has been fully specified, so that the resulting proofs can be checked by purely mechanical procedures" (Avigad, 2006) (see also for more recent work on such 'conceptual yet precise' proofs (Weber and Tanswell, 2022; Avigad, 2024)). A promising approach to this topic may thus rest on formalizing the reasoning steps and strategies used in a proof. Combined with formalizations of the more philosophical aspects of proofs (such as proof ideals), such a strategy might result in an comprehensive bridge between informal and formal proofs.

**The relation between proof theory and model theory.** Finally, we note that the investigation of semantic pollution suggests that more nuance may be given in the different ways that proof theory and model theory relate to each other, aside from soundness and completeness results. Whereas a soundness and completeness result relates a whole syntactic language to a semantics, we may take our formalization of semantic pollution as advocating more subtle distinctions between the way the object language relates to it, and a proof-theoretic language. In particular, it advocates more refinement in the role of the model-theoretic interpretation function, and whether or not it lets the interpretation of proof-theoretic syntactic elements depend on the object language interpretation.

Taking this to a different level, we note that it might be interesting to more closely investigate not only when a syntactic element violates invariance results under model equivalences in general — but when a syntactic element is 'biased' towards a particular semantics, more specifically. For instance, including labeled formulas in a proof-theoretic languages forces one to make sense of them semantically, and Kripke semantics provides the intended way of doing this. But how does the model-theoretic interpretation of labeled formulas change when adopting a different type of semantics? That is, the modal language displays flexibility with respect to many types of semantics (such as Kripke semantics, neighborhood semantics, topological semantics and algebraic semantics). It seems natural to think

that adding proof-theoretic formulas like labels makes the language less flexible in this respect. While, for instance, bilateral syntax (see e.g. (Rumfitt, 1997; Smiley, 1996)) can help enforce categoricity results for a semantics, making the relation between a proof theory and model theory more stable and complete — perhaps to a lesser extent there is also an interesting way of saying that semantically polluted syntax fixes, or rather prefers, the type of semantics that it was inspired by.

# Bibliography


M. Antonutti Marfori. Informal proofs and mathematical rigour. *Studia Logica*, 96: 261–272, 2010.

C. Antos and M. Colyvan. Explanation in descriptive set theory. In K. Robertson and A. Wilson, editors, *Levels of Explanation*. Oxford University Press, 2024.

A. Arana. On formally measuring and eliminating extraneous notions in proofs. *Philosophia Mathematica*, 17(2):189–207, 2009.

A. Arana. On the alleged simplicity of impure proof. In *Simplicity: ideals of practice in mathematics and the arts*, pages 205–226. Springer, 2017.

A. Arana. Purity and explanation: Essentially linked? In *Mathematical Knowledge, Objects and Applications: Essays in Memory of Mark Steiner*, pages 25–39. Springer, 2022.

A. Arana and M. Detlefsen. Purity of methods. *Philosophers' Imprint*, 11(2):1–20, 2011.

A. Arana and P. Mancosu. On the relationship between plane and solid geometry. *The Review of Symbolic Logic*, 5(2):294–353, 2012.

Aristotle. *Physics*. Oxford University Press, Oxford, 1954.

J. Avigad. Mathematical method and proof. *Synthese*, 153:105–159, 2006.

J. Avigad. Reliability of mathematical inference. *Synthese*, 198(8):7377–7399, 2021.

J. Avigad. The design of mathematical language. In *Handbook of the History and Philosophy of Mathematical Practice*, pages 3151–3189. Springer, 2024.

A. Avron. A constructive analysis of RM. *The Journal of symbolic logic*, 52(4): 939–951, 1987.

A. Avron. *The method of hypersequents in the proof theory of propositional non-classical logics*. na, 1996.

J. Azzouni. The derivation-indicator view of mathematical practice. *Philosophia Mathematica*, 12(2):81–106, 2004.

M. Baaz and R. Iemhoff. On skolemization in constructive theories. *The Journal of Symbolic Logic*, 73(3):969–998, 2008.

J. T. Baldwin. Formalization, primitive concepts, and purity. *The Review of Symbolic Logic*, 6(1):87–128, 2013.

D. Basin, S. Matthews, and L. Viganò. Natural deduction for non-classical logics. *Studia Logica*, 60:119–160, 1998.

M. Beeson. On the notion of equal figures in Euclid. *Beiträge zur Algebra und Geometrie/Contributions to Algebra and Geometry*, 64(3):581–625, 2023.

N. D. Belnap. Display logic. *Journal of Philosophical Logic*, pages 375–417, 1982.

K. Bimbó. *Proof theory: Sequent calculi and related formalisms*. CRC Press, 2014.

P. Blackburn. Arthur Prior and hybrid logic. *Synthese*, 150(3):329–372, 2006.

P. Blackburn, M. De Rijke, and Y. Venema. *Modal logic*, volume 53. Cambridge University Press, 2001.

P. Blackburn, J. F. van Benthem, and F. Wolter. *Handbook of modal logic*. Elsevier, 2006.

B. Bolzano. *Die drey Probleme der Rectifikation, der Complanation und der Cubierung*. Kummer, 1817.

B. Bolzano. *Theory of Science*. Oxford University Press, Oxford, 4 1837.

D. Bonnay and D. Westerståhl. Compositionality solves Carnap's problem. *Erkenntnis*, 81(4):721–739, 2016.

D. Bonnay and D. Westerståhl. Carnap's problem for modal logic. *The Review of Symbolic Logic*, pages 1–25, 2021.

G. Boolos. Iteration again. *Philosophical Topics*, 17(2):5–21, 1989.

G. Bouligand. *Structure des Théories*. Hermann, Paris, 1937.

T. Braüner. *Hybrid logic and its proof-theory*, volume 37. Springer Science & Business Media, 2010.

T. Braüner and V. de Paiva. Intuitionistic hybrid logic. *Journal of Applied Logic*, 4 (3):231–255, 2006.

C. C. Brîncuş. Inferential quantification and the $\omega$-rule. In *Perspectives on Deduction: Contemporary Studies in the Philosophy, History and Formal Theories of Deduction*, pages 345–372. Springer, 2024.

J. R. Brown. *An Introduction to the World of Proofs and Pictures*. Routledge, 1999.

K. Brünnler. Nested sequents. *arXiv preprint arXiv:1004.1845*, 2010.

R. Bull. Cut elimination for propositional dynamic logic without. *Mathematical Logic Quarterly*, 38(1):85–100, 1992.

R. Bull and K. Segerberg. Basic modal logic. In *Handbook of Philosophical Logic: Volume II: Extensions of Classical Logic*, pages 1–88. Springer, 1984.

T. Burge et al. *Truth, thought, reason: Essays on Frege*, volume 1. Oxford University Press on Demand, 2005.

J. P. Burgess. *Rigor and structure*. Oxford University Press, USA, 2015.

S. R. Buss. *Handbook of proof theory*. Elsevier, 1998.

N. A. Carlson. A connection between Furstenberg's and Euclid's proofs of the Infinitude of Primes. *The American Mathematical Monthly*, 121(5):444–444, 2014.

R. Carnap. Formalization of logic. *The Journal of Philosophy*, 40(12):332–334, 1943.

B. ten Cate. *Model theory for extended modal languages*. University of Amsterdam, 2004.

B. ten Cate and R. Koudijs. Characterising modal formulas with examples. *arXiv preprint arXiv:2206.06049*, 2022.

A. Ciabattoni, T. S. Lyon, R. Ramanayake, and A. Tiu. Display to labeled proofs and back again for tense logics. *ACM Transactions on Computational Logic (TOCL)*, 22(3):1–31, 2021.

M. Colyvan, J. Cusbert, and K. McQueen. Two flavours of mathematical explanation. *Explanation beyond causation: Philosophical perspectives on non-causal explanations*, page 231, 2018.

H. B. Curry. *A Theory of Formal Deducibility*. Notre Dame, Indiana: University of Notre Dame Press, 1950.

W. D'Alessandro. Explanation in mathematics: Proofs and practice. *Philosophy Compass*, 14(11):e12629, 2019.

W. D'Alessandro. Mathematical explanation beyond explanatory proof. *British Journal for the Philosophy of Science*, 71:581–603., 2020.

T. Dalmonte, N. Olivetti, and S. Negri. Non-normal modal logics: Bi-neighbourhood semantics and its labelled calculi. In *Advances in Modal Logic 2018*, 2018.

J. W. Dawson. Why do mathematicians re-prove theorems? *Philosophia Mathematica*, 14(3):269–286, 2006.

F. De Martin Polo. Beyond semantic pollution: Towards a practice-based philosophical analysis of labelled calculi. *Erkenntnis*, pages 1–30, 2024.

M. Detlefsen. Purity as an ideal of proof. *The philosophy of mathematical practice*, 2008.

B. Dicher. Hopeful monsters: a note on multiple conclusions. *Erkenntnis*, 85(1): 77–98, 2020.

K. Došen. Logical constants as punctuation marks. *Notre Dame journal of formal logic*, 30(3):362–381, 1989.

R. Dyckhoff. Intuitionistic decision procedures since gentzen. *Advances in proof theory*, pages 245–267, 2016.

N. Francez and R. Dyckhoff. A note on harmony. *Journal of Philosophical Logic*, 41:613–628, 2012.

G. Frege. *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number*. Northwestern University Press, Northwest, 1980.

R. French. Notational variance and its variants. *Topoi*, 38(2):321–331, 2019.

H. M. Friedman and A. Visser. When bi-interpretability implies synonymy. *Logic Group Preprint Series*, 320:1–19, 2014.

H. Furstenberg. On the infinitude of primes. *Amer. Math. Monthly*, 62(353):286, 1955.

M. Ganea. Arithmetic on semigroups. *The Journal of Symbolic Logic*, 74(1):265–278, 2009.

G. Gentzen. Untersuchungen über das logische schließen, I. *Mathematische zeitschrift*, 35, 1935.

G. Gentzen. *The collected papers*. North-Holland Publishing Company, 1969.

A. Ginammi, R. Koopman, S. Wang, S. Bloem, and A. Betti. Bolzano, Kant, and the traditional theory of concepts, a computational investigation. In P. U. Press, editor, *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*, pages 1–36. de Block, A. and Rams, G., 2020.

R. Goré. Gaggles, gentzen and galois: How to display your favourite substructural logic. *Logic Journal of the IGPL*, 6(5):669–694, 1998.

R. Goré and R. Ramanayake. Labelled tree sequents, tree hypersequents and nested (deep) sequents. In *Advances in modal logic*, pages 279–299. 2014.

I. Hacking. What is logic? *The journal of philosophy*, 76(6):285–319, 1979.

J. Hafner and P. Mancosu. The varieties of mathematical explanation. *Visualization, explanation and reasoning styles in mathematics*, pages 215–250, 2005.

M. Hallett. Reflections on the purity of method in hilbert's. *Grundlagen der Geometrie*, 2007.

Y. Hamami. Mathematical rigor and proof. *The Review of Symbolic Logic*, pages 1–41, 2019.

S. O. Hansson. Formalization in philosophy. *Bulletin of Symbolic Logic*, 6(2):162–175, 2000.

I. Hipolito and R. Kahle. Discussing Hilbert's 24th problem, 2019.

O. Hjortland and S. Standefer. Inferentialism, structure, and conservativeness. In *From Rules to Meanings*, pages 115–140. Routledge, 2018.

O. T. Hjortland. Speech acts, categoricity, and the meanings of logical connectives. *Notre Dame Journal of Formal Logic*, 55(4):445–467, 2014.

W. Hodges. *Model theory*. Cambridge university press, 1993.

L. Humberstone. *The connectives*. MIT Press, 2011.

R. Iemhoff. Remarks on simple proofs. *Simplicity: ideals of practice in mathematics and the arts*, pages 143–151, 2017.

L. Incurvati. *Conceptions of Set and the Foundations of Mathematics*. Cambridge University Press, 2020.

L. Incurvati and C. Nicolai. On logical and scientific strength. *Erkenntnis*, pages 1–23, 2024.

M. Inglis and A. Aberdein. Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica*, 23(1):87–109, 2015.

M. Inglis and J. P. Mejía-Ramos. Functional explanation in mathematics. *Synthese*, 198(Suppl 26):6369–6392, 2021.

D. Isaacson. Arithmetical truth and hidden higher-order concepts. In *Studies in Logic and the Foundations of Mathematics*, volume 122, pages 147–169. Elsevier, 1987.

S. Jaśkowski. On the rules of suppositions in formal logic. 1934.

G. I. Jojgov, R. P. Nederpelt, and M. Scheffer. Faithfully reflecting the structure of informal mathematical proofs into formal type theories. *Electronic Notes in Theoretical Computer Science*, 93:102–117, 2004.

R. Kahle and G. Pulcini. Towards an operational view of purity. *The Logica Yearbook*, 2017.

S. Kanger. *Provability in logic*. PhD thesis, Acta Universitatis Stockholmiensis, Stockholm Studies in Philosophy 1, Almqvist & Wiksell, Stockholm, 1957.

R. Kashima. Cut-free sequent calculi for some tense logics. *Studia Logica*, pages 119–135, 1994.

R. Kaye and T. L. Wong. On interpretations of arithmetic and set theory. *Notre Dame Journal of Formal Logic*, 48(4):497–510, 2007.

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

P. Kitcher. Bolzano's ideal of algebraic analysis. *Studies in History and Philosophy of Science Part A*, 6(3):229–269, 1975.

P. Kitcher. Explanatory unification. *Philosophy of science*, 48(4):507–531, 1981.

P. Kitcher. Explanatory unification and the causal structure of the world. In *Scientific explanation*, pages 410–505. MN: University of Minnesota Press, 1989.

I. Lakatos. What does a mathematical proof prove? In J. Worrall and G. Currie, editors, *Mathematics, Science and Epistemology*, page 61–69. Cambridge University Press, 1978.

I. Lakatos. *Proofs and refutations: The logic of mathematical discovery*. Cambridge university press, 2015.

M. Lange. Why proofs by mathematical induction are generally not explanatory. *Analysis*, 69(2):203–211, 2009.

M. Lange. Aspects of mathematical explanation: Symmetry, unity, and salience. *Philosophical Review*, 123(4):485–531, 2014.

M. Lange. Explanation, existence and natural properties in mathematics–a case study: Desargues' theorem. *Dialectica*, 69(4):435–472, 2015.

M. Lange. *Because without cause: Non-casual explanations in science and mathematics*. Oxford University Press, 2016a.

M. Lange. Explanatory proofs and beautiful proofs. *Journal of Humanistic Mathematics*, 6(1):8–51, 2016b.

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

M. Lange. Ground and explanation in mathematics. *Philosophers' Imprint*, 19, 2019.

S. Lavine. Quantification and ontology. *Synthese*, 124:1–43, 2000.

E. Lehet. Impurity in contemporary mathematics. *Notre Dame Journal of Formal Logic*, 62(1):67–82, 2021.

Leibniz. *New Essays on HUman Understanding*. Cambridge University Press, Cambridge, 1981.

H. Leitgeb. On formal and informal provability. In *New waves in philosophy of mathematics*, pages 263–299. Springer, 2009. doi: 10.1057/9780230245198_13.

C. I. Lewis. *A survey of symbolic logic*, volume 5. University of California press, 1918.

T. Lyon. On the correspondence between nested calculi and semantic systems for intuitionistic logics. *Journal of Logic and Computation*, 31(1):213–265, 2021a.

T. Lyon. Refining labelled systems for modal and constructive logics with applications. *arXiv preprint arXiv:2107.14487*, 2021b.

T. S. Lyon. Nested sequents for intuitionistic modal logics via structural refinement. In *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods*, pages 409–427. Springer, 2021c.

T. S. Lyon, A. Ciabattoni, D. Galmiche, D. Larchey-Wendling, D. Méry, N. Olivetti, and R. Ramanayake. Internal and external calculi: Ordering the jungle without being lost in translations. *arXiv preprint arXiv:2312.03426*, 2023.

P. Maddy. What do we want a foundation to do? In *Reflections on the Foundations of Mathematics*, pages 293–311. Springer, 2019.

P. Mancosu. Mathematical explanation: Problems and prospects. *Topoi*, 20(1): 97–117, 2001.

P. Mancosu and J. Hafner. Beyond unification. *The philosophy of mathematical practice*, pages 151–179, 2008.

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

S. Marin. *Modal proof theory through a focused telescope*. PhD thesis, Université Paris Saclay, 2018.

J.-P. Marquis. Stairway to heaven: the abstract method and levels of abstraction in mathematics. *The Mathematical Intelligencer*, 38(3):41–51, 2016.

R. Martinot. Towards a formal analysis of semantic pollution of proof systems. *The Logica Yearbook 2022*, 33(1):79–98, 2022.

R. Martinot. Ontological purity for formal proofs. *The Review of Symbolic Logic*, 17 (2):395–434, 2024a.

R. Martinot. A formal characterization of semantic pollution of modal proof systems. *(Submitted)*, 2024b.

R. Martinot and F. Poggiolesi. Purity and explanation: A systematic case study. *(Submitted)*, 2024.

T. McCarthy. Induction, constructivity, and grounding. *Notre Dame Journal of Formal Logic*, 62(1):83–105, 2021.

E. McMullin. Galilean idealization. *Studies in History and Philosophy of Science Part A*, 16(3):247–273, 1985.

J. Murzi. The inexpressibility of validity. *Analysis*, 74(1):65–81, 2014.

S. Negri. Proof analysis in modal logic. *Journal of Philosophical Logic*, 34(5): 507–544, 2005.

S. Negri. Proof theory for modal logic. *Philosophy Compass*, 6(8):523–538, 2011.

S. Negri. Non-normal modal logics: a challenge to proof theory. *The Logica Yearbook*, pages 125–140, 2016.

S. Negri and J. Von Plato. Cut elimination in the presence of axioms. *Bulletin of Symbolic Logic*, pages 418–435, 1998.

S. Negri and J. Von Plato. *Structural proof theory*. Cambridge University Press, 2008.

C. D. Novaes. The beauty (?) of mathematical proofs. *Advances in experimental philosophy of logic and mathematics*, page 63, 2019.

E. Nummela. No coincidence. *The Mathematical Gazette*, 71(456):147–147, 1987.

B. Pel. 'a remarkable artifice': Laplace, poisson and mathematical purity. *The Review of Symbolic Logic*, pages 1–37, 2023.

J. Peregrin. Rudolf Carnap's inferentialism. In *The Vienna Circle in Czechoslovakia*, pages 97–109. Springer, 2020.

A. Pillay. Remarks on purity of methods. *Notre Dame Journal of Formal Logic*, 62 (1):193–200, 2021.

E. Pimentel. A semantical view of proof systems. In *Logic, Language, Information, and Computation: 25th International Workshop, WoLLIC 2018, Bogota, Colombia, July 24-27, 2018, Proceedings 25*, pages 61–76. Springer, 2018.

C. Pincock. The unsolvability of the quintic: A case study in abstract mathematical explanation. *Philosopher's Imprint*, 15(3), 2015.

F. Poggiolesi. The method of tree-hypersequents for modal propositional logic. In *Towards mathematical philosophy*, pages 31–51. Springer, 2009.

F. Poggiolesi. *Gentzen calculi for modal propositional logic*, volume 32. Springer Science & Business Media, 2010.

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

F. Poggiolesi and F. Genco. Conceptual (and hence mathematical) explanation, conceptual grounding and proof. *Erkenntnis*, 88(4):1481–1507, 2023.

F. Poggiolesi and G. Restall. Interpreting and applying proof theories for modal logic. In *New waves in philosophical logic*, pages 39–62. Springer, 2012.

G. Pottinger. Uniform, cut-free formulations of T, S4 and S5. *Journal of Symbolic Logic*, 48(3):900, 1983.

G. Priest. An introduction to non-classical logic, cambridge univ, 2001.

M. Rathjen and W. Sieg. Proof Theory. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2024 edition, 2024.

Y. Rav. A critique of a formalist-mechanist version of the justification of arguments in mathematicians' proof practices. *Philosophia Mathematica*, 15(3):291–320, 2007.

S. Read. Semantic pollution and syntactic purity. *The Review of Symbolic Logic*, 8 (4):649–661, 2015.

M. D. Resnik and D. Kushner. Explanation, independence and realism in mathematics. *The British journal for the philosophy of science*, 38(2):141–158, 1987.

G. Restall. Multiple conclusions. In *Logic, methodology and philosophy of science: Proceedings of the twelfth international congress*, pages 189–205. Kings College Publications London, 2005.

I. Rumfitt. The categoricity problem and truth-value gaps. *Analysis*, 57(4):223–235, 1997.

S. Russ. A translation of bolzano's paper on the intermediate value theorem. *Historia Mathematica*, 7:156–185, 1980.

P. Ryan. Szemerédi's theorem: An exploration of impurity, explanation, and content. *The Review of Symbolic Logic*, 16:700–739., 2023.

P. J. Ryan. Szemerédi's theorem: An exploration of impurity, explanation, and content. *The Review of Symbolic Logic*, pages 1–40, 2021.

G. Sambin and S. Valentini. The modal logic of provability. the sequential approach. *Journal of Philosophical Logic*, pages 311–342, 1982.

P. Schroeder-Heister. Harmony in proof-theoretic semantics: A reductive analysis. In *Dag Prawitz on proofs and meaning*, pages 329–358. Springer, 2014.

P. Schroeder-Heister. Proof-Theoretic Semantics. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition, 2024.

S. Shapiro. *Foundations without foundationalism: A case for second-order logic*, volume 17. Clarendon Press, 1991.

S. Shapiro. *Philosophy of mathematics: Structure and ontology*. Oxford University Press, 1997.

S. Shapiro. Computability, proof, and open-texture. *Church's thesis after*, 70:420–455, 2006.

S. Shapiro. Identity, indiscernibility, and ante rem structuralism: The tale of i and-i. *Philosophia mathematica*, 16(3):285–309, 2008.

A. K. Simpson. *The proof theory and semantics of intuitionistic modal logic*. PhD thesis, University of Edinburgh, 1994.

T. Smiley. Rejection. *Analysis*, 56(1):1–9, 1996.

F. Steinberger. *Harmony and logical inferentialism*. PhD thesis, University of Cambridge, 2009.

F. Steinberger. Why conclusions should remain single. *Journal of Philosophical Logic*, 40(3):333–355, 2011.

M. Steiner. Mathematical explanation. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 34(2):135–151, 1978a.

M. Steiner. Quine and mathematical reduction. *The Southwestern Journal of Philosophy*, 9(2):133–143, 1978b.

L. Straßurger. What is a logic, and what is a proof? In *Logica Universalis: Towards a General Theory of Logic*, pages 135–152. Springer, 2007.

G. Sundholm. Hacking's logic. *The Journal of Philosophy*, 78(3):160–168, 1981.

C. Swoyer. Structural representation and surrogative reasoning. *Synthese*, 87: 449–508, 1991.

J. Tappenden. Fruitfulness as a theme in the philosophy of mathematics. *The Journal of Philosophy*, 109(1/2):204–219, 2012.

R. Thiele. Hilbert's twenty-fourth problem. *The American mathematical monthly*, 110(1):1–24, 2003.

D. A. Thorstad. *Purity of Methods*. PhD thesis, Bryn Mawr College, 2012.

H. Tong and D. Westerståhl. Carnap's problem for intuitionistic propositional logic. *arXiv preprint arXiv:2207.14705*, 2022.

L. Torre and S. Villata. An aspic-based legal argumentation framework for deontic reasoning. In *Computational Models of Argument: Proceedings of COMMA*, pages 266–421, 2014.

S. C. Tosatto, G. Boella, L. van der Torre, and S. Villata. Abstract normative systems: Semantics and proof theory. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

R. Turner. Programming languages as mathematical theories. In *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles*, pages 66–82. IGI Global, 2010.

A. Visser. An overview of interpretability logic. *Logic Group Preprint Series*, 174, 1997.

F. Waismann. Verifiability. In *How I See Philosophy*, pages 39–66. Palgrave Macmillan UK, London, 1968. doi: 10.1007/978-1-349-00102-6_2.

H. Wansing. Sequent calculi for normal modal propositional logics. *Journal of Logic and Computation*, 4(2):125–142, 1994.

C. Warmke. Modal semantics without worlds. *Philosophy Compass*, 11(11):702–715, 2016.

J. Warren. Infinite reasoning. *Philosophy and Phenomenological Research*, 103(2): 385–407, 2021.

K. Weber and F. S. Tanswell. Instructions and recipes in mathematical proofs. *Educational Studies in Mathematics*, 111(1):73–87, 2022.

A. Weir. Informal proof, formal proof, formalism. *Review of Symbolic Logic*, 9(1): 23–43, 2016.

R. Zach. Hilbert's program then and now. In *Philosophy of logic*, pages 411–447. Elsevier, 2007.

R. Zach. Hilbert's Program. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.