Submitted: 24.01.2018

Supervisors:

Prof. Dr. hab. Rafal Urbaniak, Ghent University, University of Gdańsk
Prof. Dr. Joke Meheus, Ghent University

Reading Committee:

Prof. Dr. Leon Horsten, Bristol University
Prof. Dr. Hannes Leitgeb, Ludwig-Maximilian University of Munich
Prof. Dr. Stewart Shapiro, Ohio State University
Prof. Dr. Peter Verdée, Université catholique de Louvain
Dr. Frederik Van De Putte, Ghent University

FACULTY OF ARTS
AND PHILOSOPHY

Faculteit Letteren en Wijsbegeerte

Pawel Pawlowski

Informally provable, refutable or neither. A non-deterministic approach to informal provability

Proefschrift voorgelegd tot het bekomen van de graad van
Doctor in de Wijsbegeerte

Promotoren: Prof. Dr. hab. Rafal Urbaniak en Prof. Dr. Joke Meheus

GHENT
UNIVERSITY

Acknowledgments

This thesis owes its existence to a legion of people and coincidences. I feel quite lucky and proud to have been a part of it.

First and foremost, I would like to thank my supervisors Rafal Urbaniak and Joke Meheus. Rafal, thank you for teaching me all the mystical aspects and arcana of logic. You are the person who helped me to put my first baby-steps in logic. You were mine Obi-wan and I was proudly your padawan.

I am in debt to Joke for reminding me that I should always confront my formal work with the initial philosophical intuitions and that I should try to broaden my interest whenever possible. I am also very grateful to you for helping me with all the practical things concerning my PhD. Especially, for the help with a bunch of emails in Dutch that I was constantly getting.

I'd like to thank all the members of the Centre for Logic and Philosophy of Science. Especially to Frederik Van De Putte for proofreading my drafts and for discussions that we had.

I am also thankful to all the people that I met during my research stays. In particular to my hosts: Joost Joosten, Leon Horsten, Christian Strasser, and Hannes Leitgeb. It was really a blessing and a very interesting lesson to be able to work with so many people. I am very thankful for all the discussions that I had with the members of MCMP and FSB, especially to Lavinia Picollo, Norbert Gratzl, Johannes Stern and Marianna Antonutti Marfori.

Most of all, I am in debt to my family and friends, in particular my mum Małgorzata and my sister Zuzanna: thank you all, for allowing me to have this wonderful possibility of studying philosophy and for all the support that I received and hopefully will receive in the future. I also would like to thank my cousin Piotr (who is some sort of a genius) and his soon-to-be wife Monika for all the talks and fishing trips that we had — they really helped to put my mind at ease. Also, I am in debt to Piotr's younger brother Grzegorz for all the help with a plenitude of peculiar quests that for some reason I had during my PhD. I am also grateful to my dearest friends Michał and Ola. They helped me to stay strong in my darkest moments. I really value their uplifting comments, talks and thoughtful moments that we shared during my studies.

So Long, and Thanks for All the Fish!

Contents

Chapter 1

Introduction

1.1 Broad philosophical perspective

The work done in this thesis generally speaking can be situated within the epistemology of mathematics and formal philosophy. We employ logical and analytical methods in order to formally characterize one of the central notions in the epistemology of mathematics — the notion of a proof. Proofs, roughly speaking, are taken to be acceptable means to establish and convince people (especially mathematicians) about the truth of a given mathematical claim.

The notion of a proof is philosophically important, for the way we understand it cements our views on the philosophy of mathematics and mathematics in general.

One of the most prominent examples of this influence is provided by mathematical intuitionism. According to the proponents of this view, an essential feature of an existence proof in mathematics is the ability to define or construct the object. On this view, if one wants to prove that there is a certain object that fits a given description, then the proof of this claim should involve a construction of the object whose existence we are about to prove.

It is quite well known that not all proofs in classical mathematics are like that. In some of them, in order to show that there is an object satisfying a certain description, we assume that there is no such object and infer a contradiction from this assumption. Such proofs are known as indirect proofs. They justify an existential claim without providing or specifying a witness for it.

Intuitionists claim that indirect proofs should not be admissible in mathematics. Thus, on the intuitionistic account, proving methods are limited to constructive ones. This results in a new explication of the notion of a mathematical proof, different from the classical notion of a proof.

The above difference has a major impact on what is quite often called the logic of proofs: the set of intuitively valid principles governing sentences that involve the notion of a proof. To get the idea of what a logic of a notion is, consider, for instance, what one can think about inferential principles governing the notion of truth. One may want to have principles assuring that the truth of a conjunction of two sentences is equivalent to a statement expressing that the first and the second conjuncts are both true. Similarly, for disjunction, one

would like to have a principle saying that a disjunction is true iff at least one of its disjunctions is true. For provability, obviously, the logic would be different than the one designed for truth. The question is: in what respects?

The logic of intuitionistic proofs has to be different from classical logic in order to limit ways of proving mathematical claims corresponding to the limitations put by the mathematical intuitionism. The logic appropriate for this notion of proofs is known as intuitionistic logic. It is strictly weaker than classical logic. For instance the double negation rule allowing to move between a formula and its double negated form is no longer valid. It is weakened from equivalence to the implication in one direction only (from φ to $\neg\neg\varphi$). So, it seems that even a *prima facie* simple change in the notion leads to major consequences.

In the thesis, however, we do not focus on intuitionism (which was used only as an example). Instead, we focus on the opposition between formal and informal proofs. The former are purely syntactical derivations stated within a fully developed axiomatic formal system, such as Zermelo-Fraenkel with the Axiom of Choice, in short ZFC (the common mathematical axiomatization of set theory). The relation between any two steps in a formal proof is purely syntactical. Within a formal system we have strict syntactical rules governing symbol manipulations that determine the ways we can generate next lines in a proof.

On the other hand, the informal notion of a proof is inspired by mathematical practice. An informal proof is a commonly accepted mathematical justification of a mathematical claim. On this account, commonly accepted means of proving things are broader than in the case of formal proofs, and incorporate insights based on semantics or intuition. What is also essential for informal proofs is the fact that not all the steps in a proof have to be spelled out in details. Quite often, mathematicians using informal proofs rely heavily on the experience and mathematical insight of the reader.

In the thesis, we study the relation between formal and informal proofs from a very particular angle. While formal logics of formal provability have been developed and thoroughly studied, the logics of informal provability are not yet fully developed. The ultimate aim of the thesis is to provide a characterization of the logic of informal provability and study philosophical implications thereof.

We argue that the logics of formal and informal proofs are different. On the account proposed in the thesis, informal provability is a partial notion: some mathematical claims are informally provable, some are informally refutable and the rest is neither. Incorporating this intuition into a formal framework results in a non-classical logic.

The logics developed in the thesis are not-deterministic. The values of complex formulas are not uniquely determined by the values of their components. This is motivated by the observation that the provability status of certain disjunctions and conjunctions of some claims which are neither provable nor refutable is not automatically settled by the status of their components.

Providing the logic of informal provability that is non-classical has a major impact on the debate about the nature of informal mathematics. First, since the logics of informal and formal provability are different, it seems that the former is not easily reducible to the latter. This means that the price that we have to

pay by increasing the mathematical rigor of proofs is not as innocent as initially one might suspect.

Second, once we provide a consistent characteristic of the notion of informal provability, it can be used as in an argument against the philosophers claiming that informal mathematics is inconsistent. On this view, using self-referential mathematical claims involving the notion of informal provability, it is possible to argue for the inconsistency of informal mathematics. Using the approach proposed in the thesis we formulate strong arguments against the claim.

Third, spelling out in detail what the logic of informal provability looks like sheds some light on a very complex relation between formal and informal provability. It shows quite clearly that both notions are not inter-reducible. Nevertheless, provability of a given claim in an axiomatic system about which we firmly believe is true still counts as a proper informal justification of the sentence. In that sense, formal proofs can be counted as formal approximations of informal proofs.

1.2 Methodology

We will employ the methodology quite common for this type of investigation. On one hand, we use philosophical analysis and intuition in order to identify the crucial properties of the notion of informal provability. Next, using conceptual analysis together with formal methods we construct formal systems. We will mostly use many-valued logics and non-deterministic semantics. Then, we study the systems and use them to approach philosophical conundrums related to the notion of provability.

1.3 Provability logics

Provability logics are, roughly speaking, modal logics meant to capture the formal principles of various provability operators (which apply to sentences) or predicates (which apply to sentence names).[1]

Historically, the first candidate for a provability logic was the modal logic **S4**. It contains as axioms all the substitutions of classical tautologies in the language with $\Box$,[2] and all substitutions of the schemata:

$$\begin{array}{ll} \text{(K)} & \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi) \\ \text{(M)} & \Box\varphi \to \varphi \\ \text{(4)} & \Box\varphi \to \Box\Box\varphi \end{array}$$

It is also closed under two rules of inference: *modus ponens* (from $\vdash \varphi$ and $\vdash \varphi \to \psi$ infer $\vdash \psi$), and *necessitation* (Nec): if $\vdash \varphi$, then $\vdash \Box\varphi$.[3]

The principles of **S4** seem sensible when $\Box\varphi$ is read as 'it is provable that φ'. Axiom **K** says that if an implication and its antecedent are provable, then so is its consequent. Axiom **M** assures us that whatever is provable should be true. In

[1] The introductory section is partially based on (Urbaniak and Pawlowski, 2018).

[2] Throughout this dissertation when talking about instances or substitutions we'll mean instances and substitutions in the full language of the system under consideration, unless specified otherwise.

[3] '$\vdash$' is just a symbol for 'is a theorem'.

the context of provability, this axiom is sometimes called the *reflection schema*. The third axiom **4** allows us to iterate provability. If something is provable, we can prove that it is by simply displaying the proof. The system was used in 1933 by Gödel to interpret intuitionistic propositional calculus (which is closely related to reasoning about provability). Alas, **S4** turned out to be inadequate as a tool for modeling the behavior of the *formal provability predicate* within axiomatic arithmetic. This was mostly due to the fact that (M), also (in the context of provability logics) called the *reflection schema*, while intuitively plausible, cannot be provable in a consistent sufficiently strong axiomatic arithmetic for the formal provability predicate of that arithmetic.

Section 1.4 introduces the basic technical machinery of first order arithmetic. We are mostly interested in a sketch of a construction of the standard provability predicate of Peano Arithmetic. Next, in section 1.5 we define a modal logic that at the time when it was constructed was thought to capture some notion of provability. Currently, it's clear that the modal operator of **S4** and a formal provability predicate of sufficiently strong formal theory have different properties. Section 1.6 presents the logic **GL** which is used to interpret the formal provability predicate. We describe the logic and define the well-known translation between the first order language of Peano Arithmetic and a propositional modal language of **GL**. In section 1.7 we present some attempts made to explicate the notion of informal provability directly within the arithmetical setting. We start with *Epistemic Arithmetic* where we extend the arithmetical language with an **S4** operator. The resulting theory is interesting albeit quite weak. Horsten (1994) proposed a more elaborate setting where informal provability is represented as a composition of two operators. The resulting theory is called *Modal Epistemic Arithmetic* and is also described in the section. A completely different way of representing informal provability over arithmetic is sketched in section 1.9. The system was developed by Horsten (2002). The idea behind it is straightforward. Informal provability is a predicate not an operator and we axiomatize it by weakening some of the rules to a certain subsystem of the theory. In section 1.10 we state the main aim of the thesis and we briefly elaborate on its content.

1.4 Provability predicate and Peano Arithmetic

Considerations of the formal provability predicate (or predicates) are usually developed in the context of an axiomatic arithmetic. This is the case for various reasons: via Gödel coding, instead of expressions, we can talk about numbers, standard arithmetical theories are usually strong enough to include a sufficiently rich theory of syntax (*modulo* coding), and arithmetic in general is a field where many results are already known and can be borrowed and applied to syntax.

For the sake of simplicity, we'll focus on one fairly standard axiomatic arithmetic: Peano Arithmetic (**PA**), although many results apply to other arithmetical theories, including some weaker ones (see for example Hájek and Pudlak, 1993, for details). The language of **PA**, $\mathcal{L}_{\mathbf{PA}}$, is a first order language with identity and a few specific symbols: $0, S, \times$ and $+$ (in the standard model of arithmetic $\mathbb{N}$ interpreted as referring to the number zero, the successor function, multiplication, and addition, respectively). For any number m, the *standard nu-*

meral for m has the form $\underbrace{S \ldots S}_{m} 0$ and is abbreviated by $\overline{m}$. The specific axioms of **PA** consist of:

$$\forall x\,(0 \neq Sx) \tag{PA 1}$$
$$\forall x, y\,(Sx = Sy \rightarrow x = y) \tag{PA 2}$$
$$\forall x\,(x + 0 = x) \tag{PA 3}$$
$$\forall x, y\,(x + Sy = S(x + y)) \tag{PA 4}$$
$$\forall x\,(x \times 0 = 0) \tag{PA 5}$$
$$\forall x, y\,(x \times Sy = (x \times y) + x) \tag{PA 6}$$

and all the instances of the induction schema:

$$\varphi(0) \wedge \forall x\,(\varphi(x) \rightarrow \varphi(S(x))) \rightarrow \forall x\,\varphi(x) \tag{PA Ind}$$

Formulas of $\mathcal{L}_{\mathbf{PA}}$ can be classified according to their logical complexity. If t is a term not containing x, $\forall x \leq t\,\varphi(x)$ and $\exists x \leq t\,\varphi(x)$ abbreviate $\forall x\,(x \leq t \rightarrow \varphi(x))$ and $\exists x\,(x \leq t \wedge \varphi(x))$ respectively. Such occurrences of quantifiers are called *bounded*, and formulas all of whose all quantifiers are bounded are called Δ_0-formulas. The hierarchy proceeds in two "layers", that of Π_n and that of Σ_n formulas. $\Pi_0 = \Sigma_0 = \Delta_0$. Σ_{n+1}-formulas are of the form $\exists x_1, \ldots, x_k\,\varphi(x_1, \ldots, x_k)$, where $\varphi(x_1, \ldots, x_k)$ is Π_n. Π_{n+1}-formulas are of the form $\forall x_1, \ldots, x_k\,\varphi(x_1, \ldots, x_k)$, where $\varphi(x_1, \ldots, x_k)$ is Σ_n. Every formula of $\mathcal{L}_{\mathbf{PA}}$ is logically equivalent to a Σ_n formula and to a Π_m formula, for some n and m (and there always exist the least such n and m).

The class of Σ_1 formulas is of particular interest, because it turns out that a function is recursively enumerable (see Smith, 2007, for a nice introduction to the topic) just in case it is Σ_1-definable. This result, for instance, makes sure that an axiomatic system which is strong enough to handle Σ_1-sentences (in a sense to be specified) is strong enough to properly handle computable functions, including those related to syntactic manipulations, and so is strong enough to prove things about syntax of a formal language within it.

We say that an arithmetical theory $\mathbf{T}$ is Σ_1-sound just in case for any Σ_1-formula φ, if $\mathbf{T} \vdash \varphi$, then $\mathbb{N} \models \varphi$ (that is, φ is true in the standard model of arithmetic). The dual notion is that of Σ_1-*completeness*. $\mathbf{T}$ is Σ_1-complete just in case for any sentence $\varphi \in \Sigma_1$, if $\mathbb{N} \models \varphi$, then $\mathbf{T} \vdash \varphi$. Interestingly, we have:

Fact 1. PA *is* Σ_1*-complete.*

There are various ways of *coding syntax*, effectively mapping syntactic objects, such as expressions, formulas, sentences and sequences thereof to natural numbers, so that each syntactic object τ of $\mathcal{L}_{\mathbf{PA}}$ is represented by its Gödel code $\ulcorner \tau \urcorner$. The details are unimportant here, so let's just focus on one coding and work with it (again, see Smith, 2007, for an accessible introduction).

Consider now any theory $\mathbf{T}$ in $\mathcal{L}_{\mathbf{PA}}$ extending **PA**. It is said to be *elementary presented* just in case there is an arithmetical Δ_0-formula $\mathtt{Ax}_{\mathbf{T}}(x)$ true of a natural number just in case it is a code of an axiom of $\mathbf{T}$. Such a formula can be further used in a fairly standard way to construct a Δ_0 arithmetical formula $\mathtt{Prf}_{\mathbf{T}}(y, x)$

which is the standard binary proof predicate of $\mathbf{T}$ such that it is true of natural numbers m and n just in case m is the code of a sequence of formulas which is a proof of the formula whose code is n (the details of the construction are inessential here). Moreover:

(Binumeration) If $\mathbb{N} \models \mathtt{Prf}_{\mathbf{T}}(m, n)$, then $\mathbf{PA} \vdash \mathtt{Prf}_{\mathbf{T}}(\overline{m}, \overline{n})$.
If $\mathbb{N} \models \neg\mathtt{Prf}_{\mathbf{T}}(m, n)$, then $\mathbf{PA} \vdash \neg\mathtt{Prf}_{\mathbf{T}}(\overline{m}, \overline{n})$.

$\mathtt{Prf}_{\mathbf{T}}(y, x)$ can be further used to define the so-called *standard provability predicate* (since we won't be talking about non-standard provability predicates, we'll simply talk about provability predicates, assuming they're standard) and the *consistency statement*:

$$\mathtt{Prov}_{\mathbf{T}}(x) := \exists y\, \mathtt{Prf}_{\mathbf{T}}(y, x)$$
$$\mathtt{Con}(\mathrm{T}) := \neg\mathtt{Prov}_{\mathbf{T}}(\ulcorner \bot \urcorner)$$

$\mathtt{Prov}_{\mathbf{T}}(y)$ is obtained from a Δ_0 formula by preceding it with an existential quantifier, and so, it is a Σ_1-formula. Therefore, by Σ_1-completeness, the first half of (Binumeration) holds for it (and the second one fails, for somewhat more complicated reasons):

If in the standard model $\mathtt{Prov}_{\mathbf{T}}(n)$ is true, then $\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{T}}(\overline{n})$.

Note however, that even though the second half of (Binumeration) fails, $\mathtt{Prov}_{\mathbf{T}}(x)$ succeeds at *defining* provability, in the sense that $\mathtt{Prov}_{\mathbf{T}}(\overline{\ulcorner \varphi \urcorner})$ is true in the standard model of arithmetic just in case in fact $\mathbf{T} \vdash \varphi$ (by the way, from now on we'll skip using the bar above numbers coding of formulas, assuming it is normally there, that is, that in the formulas we'll mention, numerals of codes of formulas are standard).

Still assuming $\mathbf{T}$ is elementary presented, $\mathtt{Prov}_{\mathbf{T}}(x)$ satisfies the following so-called *Hilbert-Bernays conditions* (Hilbert and Bernays, 1939; Löb, 1955) for any arithmetical formulas φ, ψ:

$$\mathbf{T} \vdash \varphi \text{ iff } \mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \tag{HB1}$$
$$\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \to \psi \urcorner) \to (\mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner)) \tag{HB2}$$
$$\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \mathtt{Prov}_{\mathbf{T}}(\ulcorner \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \urcorner) \tag{HB3}$$

In particular, the provability predicate of $\mathbf{T}$ can be taken to be that of $\mathbf{PA}$ itself. Also, keep in mind that most of the results apply to certain theories weaker than $\mathbf{PA}$ and to elementary presented theories extending $\mathbf{PA}$, either of which we usually choose to ignore for the sake of simplicity.

Another important piece of the puzzle will be Gödel's *incompleteness theorems*, which we include here in a somewhat modernized version:

Theorem 2. *If an elementary presented theory* $\mathbf{T}$ *extends* $\mathbf{PA}$ *and is consistent, then there is a sentence* $G \in \mathcal{L}_{\mathbf{PA}}$ *such that* $\mathbf{T} \nvdash G$ *and* $\mathbf{T} \nvdash \neg G$. *Moreover,* $\mathbf{T} \nvdash \mathtt{Con}(\mathrm{T})$.

Incompleteness follows from a more general result:

Lemma 1 (Diagonal Lemma). For any formula $\varphi(x) \in \mathcal{L}_A$ there is a sentence $\lambda \in \mathcal{L}_A$ such that

$$\mathbf{PA} \vdash \lambda \equiv \varphi(\ulcorner \lambda \urcorner).$$

The Diagonal Lemma, when we take $\varphi(x)$ to be $\neg \mathtt{Prov}_{\mathbf{PA}}(x)$, entails the existence of a sentence that can be used in the incompleteness proof, which provably satisfies the condition:

$$\mathbf{PA} \vdash G \equiv \neg \mathtt{Prov}_{\mathbf{PA}}(\ulcorner G \urcorner)$$

Such a G is independent of **PA**. The result generalizes: if a theory satisfies certain requirements and is consistent, its Gödel sentence is independent of it.

Henkin (1952) asked a related question: what happens, however, with sentences such as:

$$H \equiv \mathtt{Prov}_{\mathbf{T}}(\ulcorner H \urcorner)? \qquad \text{(Henkin)}$$

The question was soon answered by Löb (1955):

Theorem 3 (Löb)**.** *If the Diagonal Lemma applies to* **T***, and the provability predicate of a theory* **T** *satisfies (his formulation of) the Hilbert Bernays conditions (HB1-3),* $\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \varphi$ *if and only if* $\mathbf{T} \vdash \varphi$*.*

1.5 Modal logic S4

Formulas of *the language* $\mathcal{L}_M$ *of a propositional modal logic* are built from propositional variables $p_1, p_2, \ldots$, two propositional constants $\bot$ (contradiction) and $\top$ (logical truth), classical connectives $\neg, \wedge, \vee, \rightarrow, \equiv$, brackets, and unary modal connectives $\Box$ and $\Diamond$, in the standard manner. Sometimes, without loss of generality, we'll treat $\mathcal{L}_M$ as containing only a single classical connective and a single modal operator — this will shorten some definitions, and is enough to make all the other connectives definable. Given a formal language (not necessarily $\mathcal{L}_M$, the context will make the range of meta-variables clear on each occasion), we'll use lower case Greek letters $\varphi, \psi, \chi, \ldots$ as meta-variables for formulas of that language (sometimes, we'll also use σ as a metavariable for an arithmetical *sentence*).

A normal modal logic contains as axioms all the substitutions of formulas of $\mathcal{L}_M$ for propositional variables in classical tautologies, all substitutions (in $\mathcal{L}_M$) of the schema:

$$\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi) \qquad \text{(K)}$$

and is closed under two rules of inference: *modus ponens* (from $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ infer $\vdash \psi$) and *necessitation* (Nec): if $\vdash \varphi$, then $\vdash \Box\varphi$ ($\vdash \varphi$ is just a shorthand for 'φ is a theorem'). The weakest normal modal logic is called **K**, all other normal logics are its extensions.

The standard semantics of $\mathcal{L}_M$ involves *relational models* (also called *Kripke models).* A *frame* is a tuple $\mathcal{F} = \langle \mathcal{W}, \mathcal{R} \rangle$, where $\mathcal{W}$ is a non-empty set of possible worlds (or simply nodes, if you're not too much into bloated terminology) and $\mathcal{R}$ is

a binary relation on $\mathcal{W}$ ('is a possible world from the perspective of'), often called an *accessibility relation*. A model over $\mathcal{F} = \langle \mathcal{W}, \mathcal{R} \rangle$ is a triple $\mathbf{M} = \langle \mathcal{W}, \mathcal{R}, \Vdash \rangle$, where $\Vdash$ is a *forcing* (or *satisfaction*) relation between $\mathcal{W}$ and the formulas of $\mathcal{L}_M$ (think about it as 'being true in'), satisfying the following conditions for any $w \in \mathcal{W}$ and any $\varphi, \psi \in \mathcal{L}_M$:

$$\begin{array}{ll} w \nVdash \bot & w \Vdash \top \\ w \Vdash (\varphi \to \psi) & \text{iff } w \nVdash \varphi \text{ or } w \Vdash \psi \\ w \Vdash \Box\varphi & \text{iff for all } w' \in \mathcal{W}, \text{ if } w\mathcal{R}w', \text{ then } w' \Vdash \varphi \end{array}$$

It turns out that the class of formulas forced in every node in every frame is exactly the class of theorems of **K**. Sound and complete semantics for various other normal modal logics is obtained by putting further conditions on $\mathcal{R}$.

One modal logic that will be of particular interest for us is **S4**, which (in one of the formulations) is obtained from **K** by adding as axioms all the instances of the following schemata:

$$\Box\varphi \to \varphi \tag{M}$$

$$\Box\varphi \to \Box\Box\varphi \tag{4}$$

(M) is sometimes called (T), but in what follows we'll often use **T** as a variable for an axiomatic theory, so to avoid confusion, we'll stick to (M). **S4** is sound and complete with respect to frames in which the accessibility relation is reflexive ($\forall w \in \mathcal{W}\, w\mathcal{R}w$) and transitive ($\forall w_1, w_2, w_3 \in \mathcal{W}(w_1\mathcal{R}w_2 \wedge w_2\mathcal{R}w_3 \to w_1\mathcal{R}w_3)$).

Modal connectives of various modal systems admit various interpretations. $\Box$ can be interpreted as logical necessity, metaphysical necessity, physical necessity, moral obligation, knowledge, etc.. Different modal systems are taken to capture principles essential for these various notions. In what follows, we'll be concerned with the reading on which $\Box\varphi$ means 'it is provable that φ' (this reading will need further specifications, as it will turn out). Now the question is: which modal logic captures adequately the formal principles that hold for this reading?

Prima facie, **S4** seems like a decent candidate. (K) holds, because the consequent of a provable implication whose antecedent is provable is also provable. (M) holds, because whatever is provable is true. (4) holds, because if φ is provable, then by producing a proof of φ, by the same token, you are proving that it is provable (necessitation is reliable for pretty much the same reason). But are these considerations satisfactory? Not completely. First of all, we still don't know if there aren't any principles that hold for provability but are not provable in **S4**, because the argument so far was about the soundness of **S4** with respect to our intuitions about provability, not about completeness. Secondly, the argument is somewhat handwavy — it would be good to have a more precise explication of the notion of provability involved. Thirdly, even with such an explication in hand, we have to double-check if all principles of **S4** hold with respect to this explication. Things will turn out to be more complicated than one might initially expect.

Coming back to the question of whether $\Box$ of **S4** can be sensibly interpreted as the formal provability predicate: what happens when we take $\Box\varphi$ to mean $\mathtt{Prov}_{\mathbf{T}}(\ulcorner\varphi\urcorner)$? As it turns out, things fall apart quite quickly. For the sake

of simplicity we'll take the case where $\mathbf{T} = \mathbf{PA}$, but the point generalizes to consistent recursively axiomatizable extensions of **PA**.

Since $\mathbf{S4} \vdash \Box\varphi \to \varphi$ for any φ, the interpretation would require that for all $\varphi \in \mathcal{L}_{\mathbf{PA}}$, $\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner) \to \varphi$. But this, jointly with Löb's theorem, would entail that for any $\varphi \in \mathcal{L}_{\mathbf{PA}}, \mathbf{PA} \vdash \varphi$. So, if **PA** is consistent, **S4** is not the logic of the formal provability predicate of **PA**.

There is a somewhat different way to notice the inadequacy of **S4** in this context, already brought up by Feferman et al. (1986). The formula expressing $\mathtt{Con}(\mathrm{PA})$ is $\neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner\bot\urcorner)$, which is logically equivalent to $\mathtt{Prov}_{\mathbf{PA}}(\ulcorner\bot\urcorner) \to \bot$. At the modal level, this is just an axiom of **S4**, since $\Box\bot \to \bot$ falls under schema (M). Thus, if **S4** was adequate, we would have $\mathbf{PA} \vdash \mathtt{Con}(\mathrm{PA})$, which would contradict Gödel's second incompleteness theorem. Moreover, necessitation would yield $\Box(\Box\bot \to \bot)$, and so in **S4** we would be able to derive the claim that the consistency claim is derivable, which again, contradicts Gödel's second incompleteness theorem.

1.6 Modal logic GL

Let's fix our attention on the standard first order axiomatic arithmetic called *Peano Arithmetic* (**PA**). With this system in the background, instead of talking about an arithmetical formula φ, we can use a coding to represent it by some natural number, denoted by $\ulcorner\varphi\urcorner$. Once we've done this, there is (a standard way to construct) an arithmetical formula $\mathtt{Prov}_{\mathbf{PA}}(x)$ true in the standard model exactly about the codes of those formulas, which are provable in **PA**. This is the formal provability predicate of **PA**.

One crucial property of this predicate is stated by *Löb's Theorem*, according to which for any arithmetical φ, if $\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner) \to \varphi$, then already $\mathbf{PA} \vdash \varphi$. It turns out that the modal logic of provability principles of formal arithmetical provability is *Gödel-Löb logic* **GL**. Its axioms are all the substitutions of classical tautologies, all the substitutions of (K), all the substitutions of:

$$\text{(Löb)} \quad \Box(\Box\varphi \to \varphi) \to \Box\varphi$$

and the rules are *modus ponens* and necessitation. Various modal logics similar to **GL** have been developed for various notions of provability related to the standard formal provability. In the language of **GL** we can express claims such as 'p is provable', but we cannot express things such as 't is a proof of p'.[4]

Note that while (Nec) is a rule of **GL**, we cannot have $\varphi \to \Box\varphi$ as an axiom schema. While (Nec) is well-motivated (it says, in the intended interpretation, that any theorem is provably provable), the implication would say that anything *true* is provable, and that is far from obvious. In the arithmetical setting, we already have the formalized version of the second incompleteness theorem:

$$\mathbf{PA} \vdash \mathtt{Con}(\mathbf{PA}) \to \neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner\mathtt{Con}(\mathbf{PA})\urcorner)$$

[4]That is, we cannot express *explicit provability statements*. The latter task can be achieved in the so-called *Logic of Proofs* (**LP**). For a very interesting survey on this topic see (Artemov, 1994, 1998).

so if we also had:

$$\mathbf{PA} \vdash \mathtt{Con}(\mathbf{PA}) \rightarrow \mathtt{Prov}_{\mathbf{PA}}(\ulcorner \mathtt{Con}(\mathbf{PA}) \urcorner)$$

it would follow that $\mathbf{PA} \vdash \neg\mathtt{Con}(\mathbf{PA})$.

Now, is **GL** at least sound with respect to the formal provability interpretation? Well, the necessitation rule is the modal version of (HB1) and (K) is the modal version of (HB2). We can also prove in GL the modal version of (HB3), that is, $\mathbf{GL} \vdash (4)$, and so it can also be dropped when moving from **S4** to **GL**.

Fact 4. $\mathbf{GL} \vdash (4)$*, that is* $\mathbf{GL} \vdash \Box\varphi \rightarrow \Box\Box\varphi$*.*

We know that (4) is derivable in **GL**. But since (M) was the source of the problems, we also need to make sure it is not a derivable theorem schema for **GL**. Simply dropping it from the axiom schemata is not enough.

Fact 5. *It is not the case that for any* φ*,* $\mathbf{GL} \vdash \Box\varphi \rightarrow \varphi$*.*

Proof. The general structure of the argument is this. We show that all theorems of **GL** have a certain property, which $\Box p \rightarrow p$ doesn't have, and so $\Box p \rightarrow p$ is not a theorem of **GL**. The property is: *being a classical propositional tautology under the following translation.* So now we need to define a translation t from $\mathcal{L}_M$ into the classical propositional language, which translates all theorems of **GL** into classical tautologies, but at the same time translates $\Box p \rightarrow p$ into a formula whose negation is classically satisfiable. Let's start with the translation:

$$\begin{aligned} t(\bot) &= \bot \\ t(p) &= p \text{ (for all propositional variables)} \\ t(\varphi \rightarrow \psi) &= (\varphi)^\star \rightarrow (\psi)^\star \\ t(\Box\varphi) &= \top \end{aligned}$$

1. If φ is a substitution of a classical tautology, $t(\varphi)$ is a tautology. This is because the translation effectively is a substitution, and it gives a formula in the classical propositional language, in which all substitutions of tautologies are classical tautologies.

2. Let's translate the first modal axiom of **GL**. $t(\mathrm{K})$ is $\top \rightarrow (\top \rightarrow \top)$, which is a classical tautology.

3. Let's translate (Löb). $t(\text{Löb})$ is $\top \rightarrow \top$, which also is a tautology.

So we handled the axioms of **GL**, making sure their translations are classical tautologies. Now we need to take care of the inference rules.

4. Consider *modus ponens* (arguments for any classical propositional rule are pretty much the same). One can still apply *modus ponens* to $t(\varphi)$ and $t(\varphi \rightarrow \psi) = (t(\varphi) \rightarrow t(\psi))$. So if $\mathbf{GL} \vdash \varphi, \mathbf{GL} \vdash \varphi \rightarrow \psi$, we know that $\mathbf{GL} \vdash \psi$, and that the following are tautologies: $t(\varphi)$, $t(\varphi) \rightarrow t(\psi)$, and $t(\psi)$.

5. What about necessitation? Say $\mathbf{GL} \vdash \varphi$ so that also $\mathbf{GL} \vdash \Box\varphi$. Quite trivially $t(\Box\varphi) = \top$, which is a tautology.

Together, points 1-5 show that all theorems of **GL** translate into classical tautologies. Finally, we have to show that the translations of the formulas that we're interested in aren't tautologies.

6. $t(\Box p \to p) = \top \to p$, which is not a tautology.

Point 6 means that this formula is not a theorem of **GL**, which completes the proof. □

We know **S4** turned out inadequate with respect to the formal provability predicate. It turns out that **GL** does a much better job. To elaborate, we first need to explain the relation between $\mathcal{L}_M$ and $\mathcal{L}_{\mathbf{PA}}$ that will underlie what follows.

A mapping from propositional variables of $\mathcal{L}_M$ to the set of sentences of $\mathcal{L}_{\mathbf{PA}}$ is called an *arithmetical realization*. In a sense, an arithmetical realization tells us which variables are to be interpreted as which sentences of arithmetic. Given an elementary presented theory **T**, any arithmetical realization r can be extended to a **T**-interpretation $r_{\mathbf{T}}(\varphi)$ of a modal formula, by the following conditions:

$$\begin{aligned} r_{\mathbf{T}}(\bot) &= \bot \;\; r_{\mathbf{T}}(\top) = \top \\ r_{\mathbf{T}}(p) &= r(p) \text{ for any variable } p \\ r_{\mathbf{T}}(\varphi \to \psi) &= r_{\mathbf{T}}(\varphi) \to r_{\mathbf{T}}(\psi) \\ r_{\mathbf{T}}(\Box\varphi) &= \mathtt{Prov}_{\mathbf{T}}(\ulcorner r_{\mathbf{T}}(\varphi) \urcorner) \end{aligned}$$

(If you worry that $\mathcal{L}_{\mathbf{PA}}$ doesn't really contain $\bot$ and $\top$, feel free to replace them with any $\mathcal{L}_{\mathbf{PA}}$-formulas that are, respectively, refutable and provable by pure logic.) Let's call the set of all possible **T**-interpretations of $\varphi \in \mathcal{L}_M$ (under all possible realizations) $\varphi_{\mathbf{T}}$.

Given the correlation between the axioms and rules of **GL** and the Hilbert-Bernay's conditions and Löb's theorem, the adequacy of **GL** at least in one direction is clear:

Fact 6. **GL** *is sound with respect to the arithmetical interpretation, that is:*

$$\textit{If } \mathbf{GL} \vdash \varphi, \textit{ then } \mathbf{PA} \vdash \varphi_{\mathbf{T}}.$$

(where by $\mathbf{PA} \vdash \varphi_{\mathbf{T}}$ *we mean that* **PA** *proves all the members of* $\varphi_{\mathbf{T}}$*).*

In fact, implication in the opposite direction also holds, provided that **T** is Σ_1-sound, so that the claim can be strengthened to equivalence (Solovay, 1976):

Theorem 7 (Solovay Completeness)**.** *If* **T** *is* Σ_1*-sound, then for any* $\varphi \in \mathcal{L}_M$*:*

$$\mathbf{GL} \vdash \varphi \;\; \textit{if and only if} \;\; \mathbf{T} \vdash \varphi_{\mathbf{T}}.$$

This shows that given a sensible arithmetical theory, those principles of its formal provability predicate that are provable in arithmetic are adequately axiomatized by **GL**. The proof lies beyond the scope of this introduction.

1.6.0.1 Relational semantics for GL

We have drawn a connection between **GL** and the formal provability predicate. What about relational semantics for **GL**, though? As proved by Segerberg (1971), there is a natural class of relational models with respect to which **GL** is sound and complete.

Theorem 8. GL *is sound and complete with respect to the class of finite frames in which R is transitive and irreflexive.*

There is also a somewhat different class of frames with respect to which **GL** is sound and complete. We say that the accessibility relation R is *reversely well-founded* in W just in case every non-empty subset X of W has an R-maximal element (that is, a $w \in X$ such that $\neg\exists w' \in W\, wRw'$).

Theorem 9. GL *is sound and complete with respect to transitive and reversely well-founded frames.*

Notice that there is a connection between these two. Any reversely well-founded R is irreflexive, and a transitive R on a finite set is reversely well-founded just in case it is irreflexive. The result can be strengthened:

Theorem 10. GL *is sound and complete with respect to finite transitive and reversely well-founded frames.*

Since the proof employs a construction that given a formula to be checked gives an upper limit on the finite size of models to be checked, the proof by the same token proves the decidability of **GL**.

The full proof of weak completeness (that is, the one that applies to theoremhood, read on for details) is beyond the scope of this thesis. To give you a taste, however, we'll run the following interesting part of the argument to the effect that validity of (Löb) in a frame, implies that the accessibility relation is reversely well-founded. We'll argue by contraposition, by showing that if a frame isn't reversely well-founded, there is a possible world in it and a forcing relation over it, such that (Löb) fails there.

So assume R is not reversely well-founded. This means there is a set $X \subseteq W$ such that the elements of X constitute an infinite chain $w_1Rw_2Rw_3\ldots$. Take $\Vdash$ such that $w \Vdash p$ for all $w \in W \setminus X$ and $w' \nVdash \neg p$ for all $w' \in X$. Pick an arbitrary $w \in X$. Now we want to show that the antecedent of (Löb), $\Box(\Box p \to p)$, holds in w. This requires showing that $\Box p \to p$ holds in any world accessible from w. So assume wRv. We'll want to show $v \Vdash \Box p \to p$.

Either $v \in X$ or $v \notin X$. If the former, then v can access at least one world in the infinite chain. So for some $u \in X$, vRu. Since p is false in all elements of X we have $u \nVdash p$ and so $v \Vdash \Diamond\neg p$, that is $v \Vdash \neg\Box p$. But this classically entails $v \Vdash \Box p \to p$. If the latter, $v \Vdash p$, and classically $v \Vdash \Box p \to p$.

Either way, if wRv, $v \Vdash \Box p \to p$. Since our choice of v was arbitrary, and the only assumption was that wRv, this means that $w \Vdash \Box(\Box p \to p)$. This is the antecedent of (an instance) of (Löb). On the other hand, w is in a chain in X, and so it can access a world where p fails, and so $w \Vdash \neg\Box p$, which is the negation of (our instance of) (Löb).

1.7 Epistemic Arithmetic and its variants

The picture so far is that it's possible to study properties of formal provability directly in **PA** or indirectly via translation from the modal language. For informal provability the situation is a bit different. Even if we agree that **S4** is an interesting modal logic of informal provability we still do not have means of directly expressing informal provability in an arithmetical setting. In this section we will fill this hole in by introducing a couple of systems that enable one to study informal provability in arithmetic.

Historically, the first theory of informal provability is Shapiro's *Epistemic Arithmetic* (**EA**) presented in (Goodman, 1984; Shapiro, 1985) and further developed by Flagg and Friedman (1986). The idea here is to extend the standard arithmetical language $\mathcal{L}_{\mathbf{PA}}$ to $\mathcal{L}_K$ by adding a unary operator K that applies to formulas. The underlying arithmetical theory is **PA**, and the behavior of K is characterized by the following rules:

KI If $\Gamma \vdash \varphi$ and every element of Γ is epistemic, then $\Gamma \vdash K(\varphi)$
KE $K(\varphi) \vdash \varphi$

where a formula φ is ontic iff it does not contain any occurrences of the operator K and is epistemic iff it has the form $K(\psi)$ for some formula ψ. So EA has all axioms of **PA** and the above two rules for K. Note that the above rules imply **S4** principles for K.

Unfortunately, the internal logic of **EA** (that is, what in **EA** is provably provable) is quite a weak theory — in a sense, it is an elementary extension of intuitionistic Heyting Arithmetic (**HA**). Define a translation V from $\mathcal{L}_{\mathbf{HA}}$, the language of **HA**, into $\mathcal{L}_K$. We use $\bar{\varphi}$ to indicate that φ belongs to $\mathcal{L}_{\mathbf{HA}}$ as follows:

1. For atomic formulas: $V(\bar{\varphi}) = K(\bar{\varphi})$,
2. $V(\overline{\varphi \wedge \psi}) = K(V(\overline{\varphi})) \wedge K(V(\overline{\psi}))$,
3. $V(\overline{\varphi \vee \psi}) = K(V(\overline{\varphi})) \vee K(V(\overline{\psi}))$,
4. $V(\overline{\varphi \to \psi}) = K(K(V(\overline{\varphi})) \to K(V(\overline{\psi})))$,
5. $V(\overline{\varphi \equiv \psi}) = K(K(V(\overline{\varphi})) \equiv K(V(\overline{\psi})))$,
6. $V(\overline{\neg\varphi}) = K(\neg K(V(\overline{\varphi})))$,
7. $V(\overline{\forall x\, \varphi(x)}) = K(\forall x\, V(\overline{\varphi(x)}))$,
8. $V(\overline{\exists x\, \varphi(x)}) = \exists x\, KV(\overline{\varphi(x)})$.

Just for the sake of simplicity we will write φ instead of $\bar{\varphi}$ whenever it does not lead to confusion. The above translation is sound and complete in the following sense:

Theorem 11. *For every $\varphi \in \mathcal{L}_{\mathbf{HA}}$, if $\mathbf{HA} \vdash \varphi$, then $\mathbf{EA} \vdash V(\varphi)$.*

Theorem 12 (Goodman 1984)**.** *For every $\varphi \in \mathcal{L}_{HA}$, if $\mathbf{EA} \vdash V(\varphi)$,then $\mathbf{HA} \vdash \varphi$.*

EA, however, does have some interesting properties — we'll mention only two of them. The *numerical existence property* is that for any formula φ, if $\mathbf{EA} \vdash \exists x\, K\varphi(x)$ then for some natural number n, $\mathbf{EA} \vdash K\varphi(n)$. The *disjunction property* is that if $\mathbf{EA} \vdash K(\varphi \vee \psi)$ then either $\mathbf{EA} \vdash K(\varphi)$ or $\mathbf{EA} \vdash K(\psi)$.

In Shapiro's **EA**, K is a primitive operator which cannot be further analyzed. Horsten (1994) suggests that the provability operator is not primitive but complex. He distinguishes between two components of informal provability: the modal and the epistemic.

The modal component is associated with possibility. The epistemic component is explained in terms of a mathematical proof. Instead of just one operator K we have two unary operators applying to formulas: $\Diamond$ and P, where $\Diamond$ is interpreted as possibility and P intuitively stands for "some mathematician has a proof that...". In $\mathcal{L}_{\mathbf{PA}}$ extended with these two operators, $\mathcal{L}_{\mathbf{MEA}}$, and following these intuitions we present the so-called Modal Epistemic Arithmetic (**MEA**) (Horsten, 1994). The axioms of **MEA** are as follows:

1. all the axioms of **PA** with induction for the extended language,
2. $\Diamond\varphi \to \varphi$ where φ is ontic i.e. $\varphi \in \mathcal{L}_{\mathbf{PA}}$,
3. $P(\varphi) \to \varphi$,
4. $P(\varphi) \to P(P(\varphi))$,
5. $(\Diamond P(\varphi) \wedge \Diamond P(\varphi \to \psi)) \to \Diamond P(\psi)$,
6. all axioms of the modal system **S5** for $\Diamond$,

and a rule of inference: if φ is a theorem, then so is $\Diamond P(\varphi)$.

Axioms 1 and 2 are some variants of the reflection principle which is provable for P for ontic sentences, and for $\Diamond$ for all sentences. It does not follow that reflection is provable for $\Diamond P$. Axioms 3 and 4 are standard axioms for provability ((HB3) and (HB1)). Note that (HB3) works for the provability operator and (HB1) for $\Diamond P$. By a $\Diamond P$-formula we will mean any formula φ where all subformulas of φ of the form $P\chi$ are immediately preceded with $\Diamond$.

Observation 1.1. Let $\varphi \in \mathcal{L}_{\mathbf{MEA}}$ be a $\Diamond P$-formula. Then the following claims hold:

$$\mathbf{MEA} \vdash \Diamond P\varphi \to \varphi$$
$$\mathbf{MEA} \vdash \Diamond P\varphi \to \Diamond P\Diamond P\varphi$$

The above observation shows that we have a certain version of reflection schema and certain version of (HB3), at least for a restricted class of formulas.

1.8 Problems with informal provability understood as an operator

The main aim of treating provability as an operator is to circumvent the impossibility that arises for the formal provability predicate — that of having all HB conditions and all the instances of the reflection schema at the same time.

Theorem 13 (Montague's theorem)**.** *Peano Arithmetic, if consistent, cannot contain (or be consistently extended to contain) a (possibly complex) predicate for which all Hilbert-Bernays conditions and all instances of the reflection schema hold.*

Proof. Suppose that there is such a predicate, call it P. We will use a natural deduction system. Argue inside the theory:

1. $\lambda \equiv P(\ulcorner\neg\lambda\urcorner)$	Diagonal lemma
1.1 λ	Hypothesis
1.2 $P(\ulcorner\neg\lambda\urcorner)$	equivalence elimination: 1,1.1
1.3 $\neg\lambda$	modus ponens and reflection schema: 1.2
2. $\neg\lambda$	reductio ad absurdum: 1.1 $\rightarrow$ 1.3
3. $P(\ulcorner\neg\lambda\urcorner)$	(HB1)
4. $\neg P(\ulcorner\neg\lambda\urcorner)$	CL, 1, 2
5. contradiction	CL, 3, 4

□

In order to prove Montague's theorem one applies the diagonal lemma to a certain formula involving the provability predicate. But if provability is treated as an operator, we cannot use the Diagonal Lemma to generate this paradoxical formula.

MEA is capable of proving variants of the reflection schema. It is an interesting result, for the name of the game here is to gather as many instances of the reflection schema as possible without inconsistency. Unfortunately, the theory has some other philosophical problems:

1. The choice which rules are postulated for P and which are postulated for $\Diamond$ seems somewhat arbitrary. It is possible to consider different combination of those rules. For instance, to add axiom (K) directly for P.

2. The reflection schema is available only for $\Diamond P$. It is not clear why other types of reflection shouldn't be introduced. For instance, reflection restricted to Σ_1 formulas doesn't look completely insane.

3. Usually provability is treated as a predicate and not as an operator. There seems to be no motivation for using an operator, independent of blocking t Montague's theorem.

4. Both **EA** and **MEA** seem to be a bit too weak — there are translations to **HA** which preserve theorems.

1.9 Informal provability as a predicate — splitting solution

Another strategy is to treat informal provability as a predicate and weaken some of the Hilbert-Bernays conditions for this predicate. Again, expand $\mathcal{L}_{\mathbf{PA}}$ with an additional predicate P for informal provability, thus obtaining a new language $\mathcal{L}_P$. The idea here is straightforward: we divide the set of problematic principles (HB conditions and the reflection schema) for the additional predicate P between

two theories: **PEA** and its basis **BPEA**. Then we add to **PEA** all the instances of the axiom saying that if something is derivable in the basis, it is informally provable.

We will start with the basis of **PEA** (**BPEA**) (Horsten, 1997), which is defined by:

Basis Axiom 1	**PA** in extended language with induction extended to $\mathcal{L}_P$
Basis Axiom 2	$P(\ulcorner\varphi\urcorner) \rightarrow (P(\ulcorner\varphi \rightarrow \psi\urcorner) \rightarrow P(\ulcorner\psi\urcorner))$ for all $\varphi, \psi \in \mathcal{L}_P$
Basis Axiom 3	$P(\ulcorner\varphi\urcorner) \rightarrow P(\ulcorner P(\ulcorner\varphi\urcorner)\urcorner)$ for all $\varphi \in \mathcal{L}_P$

So, we have (K) and (4) for P. By $\texttt{Prov}_B$ we mean the standard provability predicate of **BPEA**. **PEA** is given by the following axioms:

Axiom 1	**PA** in the extended language with induction extended to $\mathcal{L}_P$
Axiom 2	$P(\ulcorner\varphi\urcorner) \rightarrow \varphi$ for all $\varphi \in \mathcal{L}_P$
Axiom 3	$\texttt{Prov}_B(\ulcorner\varphi\urcorner) \rightarrow P(\ulcorner\varphi\urcorner)$ for all $\varphi \in \mathcal{L}_P$

We have the reflection schema for P. Notice that we do not have (Nec) for P, but we have the implication $\texttt{Prov}_B(\ulcorner\varphi\urcorner) \rightarrow P(\ulcorner\varphi\urcorner)$, which together with the reflection schema gives us $\texttt{Prov}_B(\ulcorner\varphi\urcorner) \rightarrow \varphi$ which is a certain version of (Nec).

These theories are still under investigation. One of the nice things about **PEA**, apart from the reflection schema holding in it, is the fact that **PEA** has nice models.

Fact 14. PEA *has a model based on the standard model of arithmetic.*

However, it seems that the philosophical motivations underlying the system are somewhat lacking. While informal provability seems unified, this system clearly has two separate layers. The restrictions on the claims for which reflection can be used is still there — it's just that they're somewhat less visible, because they arise at the point in which a restriction is put on what can be provably provable (**Axiom 3**). Yes, **Axiom 2** guarantees that reflection is provable for any ϕ, but given that the internal logic of P is built starting from the formal provability predicate of **BPEA**, it holds universally at the price of being idle on many occasions.

1.10 The aim of the thesis

As one can already tell from the title, the main character of the thesis is the notion of informal provability. The notion itself is rather elusive and obscure. Generally, it's quite hard to provide a positive characterization of it. There are two things that we can do to partially explicate the notion and provide certain intuitions. First, we can contrast this notion with the notion of formal provability. Second, we can try to capture the intuitively valid inference patterns involving the notion. We are interested in both. Keep in mind though that the main aim is to characterize intuitively valid inference patterns involving a certain explication of informal provability. In order to do so, we pose the following research questions:

1. Is there a difference between formal and informal provability?

2. Can we provide an interesting reading of informal provability as a partial notion and provide a decent formal framework for it?

3. What are the paradoxes related to the notion of informal provability? Is it possible to avoid them in the resulting theory?

For the first question, we argue that indeed there is a difference between these two notions. The crucial difference, as far as this thesis is concerned, is the validity of the reflection schema for informal provability.

For the second question, the notion of informal provability is treated as a partial one. Thus, on this account, some mathematical claims are informally provable, some others are refutable and some are neither informally provable nor informally refutable. To describe this approach in a more technical manner, we develop a non-deterministic logic BAT and its extension CABAT. In these logics, the explication of informal provability via non-deterministic semantics is not crazy at all. *Au contraire!* This explication seems to be very close to the very common, in mathematical practice, idea that either a sentence can be proved, or disproved, or neither.

It's possible to provide a philosophically coherent and somewhat convincing story on how the indeterminism may be used. Roughly speaking, mathematical sentences can be informally proved, informally refuted or informally undecided. Yet, some complex sentences built from two informally undecided formulas are informally provable, and some other informally refutable. For instance, the disjunctions of two undecided sentences can be informally provable (The Continuum Hypothesis and its negation) or still be informally undecided (for instance a disjunction of the Continuum hypothesis with itself). This motivates the non-deterministic approach.

On this non-deterministic reading it is possible to avoid persistent paradoxes of informal provability without paying a huge bill. Most of the crucial inference patterns for informal provability are still valid in the framework. BAT and CABAT are not very easy beasts to tame, though. Some of their properties are unusual.We construct proof-systems for both logics, then take a detour into semantical aspects of the logics. Notably, the lack of neighborhood semantics for CABAT is proven. The situation is even worse — neither the theorems of CABAT nor the consequence relation of BAT can have a finitely many-valued deterministic semantics.

Proceeding with the inquiry, we take a closer look at the paradoxes of informal provability. It turns out that some of them are avoided by switching from classical logic to CABAT in the background. A version of the dialetheist argument for the inconsistency of mathematics is presented and discussed. We argue that the argument fails and that if we switch to CABAT in the background we need not to worry about it.

Lastly, the BAT framework is developed into a full first order theory. The main idea is to use three-valued structures. On this account, an interpretation of a predicate ascribes to it a triple whose elements are called the extension, the anti-extension and the fringe. Intuitively the first set corresponds to all the elements

for which the predicate holds, anti-extension is the set of all elements for which the predicate does not hold and the fringe is a set of all the elements for which the predicate is not applicable. These sets can be used to define a triple which serves as a partial satisfaction relation restricted only to atomic formulas. Thus all atomic sentences have exactly one of the values: 1,n or 0. Next, we extend this partial satisfaction relation to the full language by means of evaluations. Evaluations are total functions from the set of sentences of a given language into the set of values such that they agree with values of atomic formulas determined relative to a structure. According to this, one structure can have many different underlying evaluations. We limit our attention to the so-called BAT-evaluations. These are based on the definition of a BAT evaluation for propositional language and similarly as in the propositional case they are not unique. As for quantifiers, they are treated as "infinitary" disjunction and conjunction.

1.11 The structure of the thesis

The dissertation is paper-based. Each chapter is either an accepted paper or a draft of a paper. Chapter 2 is an accepted and forthcoming paper in the *Review of Symbolic Logic*. In the paper we briefly discuss the relation between formal and informal provability, concluding that these two notions are different. Next, we propose a non-deterministic interpretation of informal provability and develop the many-valued non-deterministic logic BAT and its extension CABAT. We study some of their technical properties, focusing on certain inferential patterns involving informal provability. The moral from the paper is that it's possible to retain quite a large set of intuitive principles of informal provability. Chapter 3 is an accepted and forthcoming paper in the *Logic Journal of the IGPL*. It's devoted solely to proof-theoretic investigations of BAT and CABAT. Inspired by Carnielli (1987) and Priest (2001) we construct tree-like proof systems for BAT and CABAT and prove strong completeness. Chapter 4 is a submitted draft. It generalizes the proof-theoretic framework of BAT and CABAT. Using the framework it's possible to generate proof-systems for any finitely-many valued deterministic and non-deterministic consequence relation. Chapter 5 is also a submitted draft. There, we are concerned with purely semantical investigations of BAT and CABAT. We prove that both logics can't be pinned down by a vast class of well-known semantics. The main result in the paper is the proof that there is no finitely many-valued deterministic semantics for CABAT. Chapter 6 is also a submitted draft. It's more philosophical — we present well-known paradoxes involving various notions of provability and how they correspond to certain limitation theorems. Dialetheist arguments to the effect that informal mathematics is inconsistent is thoroughly discussed. We dissect some of its premises and argue that the argument does not work. Next, we discuss the paradoxes and corresponding theorems with CABAT in the background. It turns out that most of the paradoxes are blocked and at the same time the set of admissible principles of informal provability in CABAT is quite large and intuitive.

Chapter 7 contains conclusions and sketches the future perspective of logics of informal provability. We start with a generalization of BAT to a first order version. To do so, we use the notion of a three-valued structure. Based on

that, we define the interpreting function which uniquely determine the values of atomic formulas. Similarly to the propositional case, the values of all atomic formulas are uniquely determined by the interpretation. To ascribe logical values to complex formulas, we introduce the notion of an evaluation. On this account, an evaluation determines logical values for all complex formulas and has to agree on the values of atomic formulas determined by the structure together with its interpretation. In particular we are interested in a certain type of evaluations which we call BAT evaluations. These evaluations are generated by applying the meaning of BAT connectives and quantifiers are treated as "infinitary" BAT conjunction and disjunction. In the last subsection we state the conclusions of the thesis.

Chapter 2

Many-valued logic of informal provability: a non-deterministic strategy[1]

Abstract

Mathematicians prove theorems in a semi-formal setting, providing what we'll call *informal proofs*. There are various philosophical reasons not to reduce informal provability to formal provability within some appropriate axiomatic theory (Leitgeb, 2009; Antonutti Marfori, 2010; Tanswell, 2015), but the main worry is that we seem committed to all instances of the so-called reflection schema: $\mathtt{B}(\varphi) \to \varphi$ (where $\mathtt{B}$ stands for the informal provability predicate). Yet, adding all its instances to any theory for which Löb's theorem for $\mathtt{B}$ holds leads to inconsistency.

Currently existing approaches (Shapiro, 1985; Horsten, 1996, 1998) to formalizing the properties of informal provability avoid contradiction at a rather high price. They either drop one of the Hilbert-Bernays conditions for the provability predicate, or use a provability operator that cannot consistently be treated as a predicate.

Inspired by (Kripke, 1975), we investigate the strategy which changes the underlying logic and treats informal provability as a partial notion. We use non-deterministic matrices to develop a three-valued logic of informal provability, which avoids some of the above mentioned problems.

Keywords. Informal provability, many-valued logic, non-deterministic semantics, Löb's theorem, paradoxes of provability

AMS classification. 03A05, 00A30

[1]This a joint paper with Rafal Urbaniak. I am the first author and the paper is accepted and forthcoming in the *Review of Symbolic Logic*.

Acknowledgments. Research on this paper has been funded by the Research Foundation Flanders (FWO) (FWO) and National Science Centre (NCN, 2016/22/E/HS1/00304). The authors would like to express their gratitude to all those who commented on the earlier versions of this paper or contributed to discussions about the topic when the material was presented: Cezary Cieśliński, Leon Horsten, Hannes Leitgeb, Frederik Van De Putte and Stanislav Speransky.

2.1 Formal vs Informal provability

In common mathematical practice mathematical claims are justified or proven in an informal way. Informal proofs are not stated in a proper formal language, but rather in a mixture of a native language expanded with mathematical notation. They abide by a different canon of rigour than formal proofs. From the perspective of fully formalized proofs, in informal proofs some inference steps seem to be missing. It is not even clear what counts as an axiom and some simple facts are said to be justified merely on the basis of mathematical insight (or intuition). Yet, the existence of an informal proof of a mathematical statement is a very good reason to take the claim to be true (or established). Provability in the above sense will be called *informal provability*.

On the other hand, there exist *formal proofs*, given in a fully formalized axiomatic theory by means of a fully specified formal proof system. Formal provability in this sense is always relative to some axiomatic theory.

The relation between formal and informal provability is often explained by the so-called *standard view*. The proponents of this view argue that any informal proof is at least in principle reducible to a proper proof in an appropriate axiomatic system (usually, ZFC). On this view, informal proofs are just sloppy, incomplete versions of formal proofs.

Yet, there are reasons to think that there is at least a conceptual difference between these notions. Some philosophers (Horsten, 2002; Leitgeb, 2009; Antonutti Marfori, 2010) argue against the standard view. According to them, the standard view does not fully explain why informal proofs are quite good at convincing mathematicians, whereas formal ones are not. They also point out that the role of axioms and definitions is quite different in both kinds of proofs and that there is no clear procedure for converting an informal proof into a formal one or for associating informal proofs with their formal counterparts (Tanswell, 2015).

For us, the most important argument for the difference between formal and informal provability lies in general principles valid for informal provability. There is an agreement that principles of formal provability are satisfied for informal provability. Yet, those principles are not enough, they do not express the reliability of informal proofs. The additional principle, which is thought to be sound for informal provability is the *reflection schema*. It roughly says that any informally provable sentence is also true.

Unfortunately, the language of any arithmetical theory T containing Peano arithmetic, cannot contain a formula for which the combination described above holds. We will elaborate on this in section 2.7.

Current theories of informal provability (Horsten, 2002) have to face the cost of adding all the instances of the reflection schema for a new informal provability predicate. It is quite high: some other principles which intuitively hold for informal provability (such as some of Hilbert-Bernays derivability conditions) have to go.

Another strategy of constructing a theory of informal provability is to pay a different price for adding all the instances of the reflection schema for informal provability. In such systems, provability can only be treated as an operator, but, under the threat of inconsistency, not as a predicate (Shapiro, 1985).

By the end of the paper the reader will notice that in the system we present all the instances of reflection for informal provability can be added, while (some variants of) other intuitive principles are preserved. Our current goal is only to discuss the propositional level of the inferential machinery, so showing that our strategy can be consistently extrapolated to the informal provability *predicate* lies beyond the scope of this paper. However, it will become clear that the reasons which blocked the move to the predicate level for other systems are not going to constitute a similar obstacle in the case at hand.

We would like to suggest an unexplored strategy out of these difficulties, which stems from the intuitions that some of the solutions proposed in Kripke's theory of truth can be used to approach provability.

Instead of dropping or restricting Hilbert-Bernays conditions we will change the underlying logic. Most notably, our goal is to explore the option of treating mathematical provability as a partial notion — after all, there is an intuitive division of mathematical claims into provable, refutable and undecidable.

In the standard Kripke construction we rely on the Strong Kleene logic to deal with the partial truth predicate. But Kleene logic is not appropriate for modelling informal provability. It seems that informal provability doesn't have a truth-functional nature. Generally it is not always the case that disjunctions of two independent sentences of a given theory are independent of that theory.

We'll limit our attention to the arithmetical setting since it is at the same time quite simple to handle and expressive enough. The logic developed in this paper is propositional and it still needs to be further developed to the full first order version. Yet, some properties of informal provability can be studied in the propositional setting, and doing so seems like a good place to start, especially as it will turn out on page 2.7 that the propositional level is where most of the action is.

2.2 The non-deterministic strategy

Let $\mathcal{L}$ be a propositional language (understood as the set of all well-formed formulas) constructed from propositional variables $W = \{p_1, p_2, \dots\}$ and Boolean connectives $(\neg, \wedge, \vee, \rightarrow, \equiv)$ in the standard manner. We will use Greek letters $\varphi, \psi, ...$ as meta-variables for formulas. The language that results from extending the set of Boolean connectives with one unary operator $\mathtt{B}$ will be denoted $\mathcal{L}_{\mathtt{B}}$. We will use $\mathtt{B}$ to express provability within the object language.

By an *assignment* we mean any function $v : W \mapsto Val$, where Val is a set of values. By an *evaluation* e_v built over an assignment v we will mean a function

assigning values to all well-formed formulas ($e_v : \mathcal{L} \mapsto Val$) agreeing with v on W (propositional variables), and satisfying some additional constraints determined by a given logic.

In the case of standard classical propositional logic, evaluations are unambiguously determined by assignments. For each assignment there is exactly one evaluation extending it.

It is possible to construct sensible logics for which this uniqueness fails. One nice example is the propositional variant of paraconsistent logic CLuN (Batens and De Clercq, 2004).[2] The standard semantics of CLuN is similar to the semantics of classical propositional logic with one difference: the truth conditions for negation are different.

Both for classical logic and for CLuN we have $Val = \{0, 1\}$. In classical propositional logic $e_v(\neg\varphi) = 1$ iff $e_v(\varphi) = 0$. In CLuN this equivalence is weakened to an implication: if $e_v(\varphi) = 0$, then $e_v(\neg\varphi) = 1$. (Clauses for the rest of connectives are the same as in classical propositional logic.) In other words, CLuN allows for gluts for negation: both φ and $\neg\varphi$ can be true in one and the same evaluation.

The standard semantics of CLuN has another interesting feature. It is non-deterministic: assignments of values to propositional variables do not uniquely determine evaluations of all formulas. One and the same assignment might be extended in different ways to different evaluations, as long as they obey classical clauses for connectives other than negation and the implication above for negation. For instance, if $v(p) = 1$, there is one evaluation e^1_v such that $e^1_v(\neg p) = 0$ and there is another one e^2_v such that $e^2_v(\neg p) = 1$.

2.3 Non-deterministic matrices for provability

We apply a similar trick to develop a non-deterministic semantics for a logic which would help us model the notion of informal provability.

The logic will be three-valued: we take the set of values $Val = \{0, n, 1\}$. The intended interpretation of the values is as follows. 1 stands for *(informally) provable*, 0 represents *(informal) refutability* and n stands for *being neither (informally) provable, nor (informally) refutable*. This is the *synchronic* interpretation, on which whether something is informally provable or refutable doesn't depend on the stage of the development of mathematics or on anyone's state of knowledge.

We will develop a logic with provability values. One might think that this approach is strictly speaking anti-realist (because the values aren't interpreted in terms of what happens in the "external world" but rather in terms of the properties of the system), but we are not deeply committed to this way of thinking about it. One can be a truth-value realist or an ontological realist while using our logic to reason about provability and at the same time being aware that provability and truth are quite different.

[2] A general framework for non-deterministic logics can be found in (Avron and Zamanski, 2011). Particular systems discussed there have, however, quite different motivation from ours, and quite different matrices.

Perhaps, one can think of these values *diachronically* by assuming that what is informally provable changes through time as new proofs are developed. In this sense, 1 would stand rather for *being informally proven*, 0 for *being informally refuted* and i for *being neither*. While we conjecture that this interpretation should abide by the same intuitively valid inferential principles, due to the limited scope of this paper we have to postpone a proper discussion of this reading aside.

Recall that $\mathcal{L}_{\mathtt{B}}$ is the propositional language with provability operator $\mathtt{B}$. We now move to specifying the semantics for connectives of $\mathcal{L}_{\mathtt{B}}$ by means of non-deterministic matrices. Let's start with negation:

- $e_v(\varphi) = 1$ iff $e_v(\neg\varphi) = 0$.
- $e_v(\varphi) = 0$ iff $e_v(\neg\varphi) = 1$.
- $e_v(\varphi) = n$ iff $e_v(\neg\varphi) = n$.

A given formula is informally provable iff its negation is informally refutable. A given formula is informally refutable iff its negation is informally provable. A formula is undetermined iff its negation is.

For disjunction we introduce non-deterministic clauses. The equivalence

$$e_v(\varphi \vee \psi) = 1 \text{ iff } e_v(\varphi) = 1 \text{ or } e_v(\psi) = 1$$

is weakened to one direction only:

$$\text{If } e_v(\varphi) = 1 \text{ or } e_v(\psi) = 1 \text{ then } e_v(\varphi \vee \psi) = 1.$$

The full set of clauses for disjunction is:

- If $e_v(\varphi) = 1$ or $e_v(\psi) = 1$, then $e_v(\varphi \vee \psi) = 1$.
- $e_v(\varphi \vee \psi) = 0$ iff $e_v(\varphi) = e_v(\psi) = 0$.
- If $e_v(\varphi) = 0$, $e_v(\psi) = n$, then $e_v(\varphi \vee \psi) = n$.
- If $e_v(\varphi) = n$, $e_v(\psi) = 0$, then $e_v(\varphi \vee \psi) = n$.
- If $e_v(\varphi) = n$, $e_v(\psi) = n$, then $e_v(\varphi \vee \psi) = n$ or $e_v(\varphi \vee \psi) = 1$.

The intention behind the introduction of non-determinism is this. We want to allow for the possibility of there being informally (absolutely) undecidable mathematical sentences (without saying that there are any). Yet, even for such sentences (if there are any), some disjunctions built from them might be informally undecidable, while some others will be informally provable. Say φ and ψ are informally undecidable (and therefore, so is $\neg\varphi$). Then, while we might think that $\varphi \vee \psi$ is informally undecidable, we might be inclined to think that $\varphi \vee \neg\varphi$ is informally provable despite φ not being decidable.

For instance, you might be inclined to think that the Continuum Hypothesis (CH) is informally undecidable, while $CH \vee \neg CH$ is still informally provable, being a logical truth. This however, clearly doesn't mean that $CH \vee CH$ is provable, and so not every disjunction of undecidable sentences is decided.

Conjunction $\varphi \wedge \psi$ is taken to have the same matrix as $\neg(\neg\varphi \vee \neg\psi)$, and so:

- If $e_v(\varphi) = 0$ or $e_v(\psi) = 0$ then $e_v(\varphi \wedge \psi) = 0$.

- $e_v(\varphi \wedge \psi) = 1$ iff $e_v(\varphi) = e_v(\psi) = 1$.
- If $e_v(\varphi) = 1$, $e_v(\psi) = n$ then $e_v(\varphi \wedge \psi) = n$.
- If $e_v(\varphi) = n$, $e_v(\psi) = 1$ then $e_v(\varphi \wedge \psi) = n$.
- If $e_v(\varphi) = n$, $e_v(\psi) = n$ then $e_v(\varphi \wedge \psi) = n$ or $e_v(\varphi \wedge \psi) = 0$.

The idea for the non-deterministic case for conjunction is as follows. For some informally undecidable sentences we may be able to prove that they are mutually contradictory, which makes their conjunction informally refutable. For some others it may be impossible, and so their conjunction remains informally undecidable.[3]

Implication is taken to have the same matrix as $(\neg\varphi \vee \psi)$, and so:[4]

- If $e_v(\varphi) = 0$ then $e_v(\varphi \to \psi) = 1$.
- $e_v(\varphi \to \psi) = 0$ iff $e_v(\varphi) = 1$ and $e_v(\psi) = 0$.
- If $e_v(\varphi) = n$, $e_v(\psi) = n$ then $e_v(\varphi \to \psi) = n$ or $e_v(\varphi \to \psi) = 1$.
- If $e_v(\varphi) = n$, $e_v(\psi) = 1$ then $e_v(\varphi \to \psi) = 1$.
- If $e_v(\varphi) = n$, $e_v(\psi) = 0$ then $e_v(\varphi \to \psi) = n$.
- If $e_v(\varphi) = 1$, $e_v(\psi) = n$ then $e_v(\varphi \to \psi) = n$.
- If $e_v(\varphi) = 1$, $e_v(\psi) = 1$ then $e_v(\varphi \to \psi) = 1$.

Equivalence has the same matrix as $((\varphi \to \psi) \wedge (\psi \to \varphi))$, and therefore:

- $e_v(\varphi \equiv \psi) = 1$ if $e_v(\varphi) = e_v(\psi) = 1$ or $e_v(\varphi) = e_v(\psi) = 0$.
- $e_v(\varphi \equiv \psi) = 0$ if ($e_v(\varphi) = 1$ and $e_v(\psi) = 0$) or ($e_v(\varphi) = 0$ and $e_v(\psi) = 1$).
- $e_v(\varphi \equiv \psi) = n$ if exactly one of ψ, φ has value n.

While this doesn't need to be stated and follows from the above, notice that if $e_v(\varphi) = e_v(\psi) = n$ then $e_v(\varphi \equiv \psi)$ is either $0, n, 1$.

The intended reading of $\mathtt{B}\varphi$ is 'φ is informally provable.' The matrix for $\mathtt{B}$ is non-deterministic:

- $e_v(\mathtt{B}\varphi) = 1$ iff $e_v(\varphi) = 1$.
- If $e_v(\mathtt{B}\varphi) = 0$, then $e_v(\varphi) = 0$ or $e_v(\varphi) = n$.
- If $e_v(\mathtt{B}\varphi) = n$, then $e_v(\varphi) = n$.

[3]Notice that just because $\varphi \wedge \psi$ has the same truth table as $\neg(\neg\varphi \vee \neg\psi)$, it doesn't follow that the substitution of expressions of this form preserves the value under an interpretation. This will fail due to indeterminacy. (The substitutability will be regained once we move from BAT to CABAT.)

[4]There are other ways to introduce implication in many-valued contexts, but given how, as it will turn out, the behavior of implication deserves additional attention, we postpone the discussion of various ways it can or cannot be introduced to another paper.

The intuition behind these conditions is the following.

If a formula is informally provable ($e_v(\varphi) = 1$), then giving its own proof is also a proof of its provability ($e_v(\mathtt{B}\varphi) = 1$), and the other way around. If a formula is informally refutable $e_v(\varphi) = 0$, then giving its own refutation is also a refutation of its provability ($e_v(\mathtt{B}\varphi) = 0$). If a formula is informally undecidable ($e_v(\varphi) = n$), then one of two things may happen. First, it may be the case that the undecidability of that formula is informally provable, and so its informal provability is refutable ($e_v(\mathtt{B}\varphi) = 0$). Second, it may be the case that its absolute informal undecidability is not informally provable, and so its absolute informal provability is informally undecidable ($e_v(\mathtt{B}\varphi) = n$).

All these conditions are captured by the following tables:

$\neg$	φ
0	1
n	n
1	0

$\vee$	0	n	1
0	0	n	1
n	n	n/1	1
1	1	1	1

$\wedge$	0	n	1
0	0	0	0
n	0	0/n	n
1	0	n	1

$\rightarrow$	0	n	1
0	1	1	1
n	n	n/1	1
1	0	n	1

$\equiv$	0	n	1
0	1	n	0
n	n	0/n/1	n
1	0	n	1

B	φ
1	1
n/0	n
0	0

Because we interpret value 1 as **B**eing an **A**bsolute **T**heorem (BAT), we call the logic thus obtained BAT and we'll use the bat symbol 🦇 to denote its consequence relation, which we define as follows.

A BAT-assignment v is a function from propositional variables W to $\{0, n, 1\}$. A BAT-evaluation over an assignment v is a function which assigns values to all formulas of $\mathcal{L}_{\mathtt{B}}$, agrees with v on W and obeys the constraints we gave for the connectives. Notice that due to non-deterministic clauses, one and the same assignment might underlie multiple evaluation functions.

By Γ🦇φ, where Γ is a set of formulas, we will mean that any BAT-evaluation which assigns 1 to all formulas in Γ assigns 1 to formula φ. We say that φ is a BAT-tautology iff $\emptyset$🦇φ. We say that φ is a countertautology of BAT iff $\emptyset$🦇$\neg\varphi$.

2.4 Properties of BAT

First, note:

Theorem 15. *BAT has neither tautologies nor countertautologies.*

Proof. Because n is contagious, it is easy to see by induction on formula complexity that for the assignment v which assigns n to all propositional variables and for any formula φ there will be a way of extending v to e_v such that $e_v(\varphi)$ will be n. □

Theorem 16. *Let $\Gamma \subseteq L$ and $\varphi \in L$, then for any set of formulas Γ and any formula φ if Γ🦇φ then $\Gamma \models \varphi$, where $\models$ is the classical consequence relation (we'll use $\models$ in this sense throughout the paper).*

Proof. By contraposition suppose that $\Gamma \not\models \varphi$. Then there is an evaluation over an assignment v such that $e_v(\psi) = 1$ for all $\psi \in \Gamma$ and $e_v(\varphi) = 0$. It is easy to see that e_v is also a BAT-evaluation. For the assignment v is classical (it is into $\{0, 1\}$) and BAT-evaluations behave in the same manner as classical evaluations over classical assignments. Hence, there is at least one BAT-assignment which makes all formulas of Γ true and φ false, which means that it is not the case that $\Gamma \vdash\!\!\!\!\!\!\diagup\; \varphi$ (that is, $\Gamma \vdash\!\!\!\!\!\!\diagup\; \varphi$). □

Quite expectedly, classical consequence is strictly stronger than BAT-consequence:

Theorem 17. *There are some $\Gamma \subseteq L$ and $\varphi \in L$ such that $\Gamma \models \varphi$ but $\Gamma \vdash\!\!\!\!\!\!\diagup\; \varphi$.*

Proof. For instance, $\neg(\varphi \wedge \psi) \vdash\!\!\!\!\!\!\diagup\; \neg\varphi \vee \neg\psi$. Take any evaluation for which

$$e_v(\varphi) = n = e_v(\psi), e_v(\varphi \wedge \psi) = 0, e_v(\neg\varphi \vee \neg\psi) = n.$$

□

The following table illustrates the assessment of some standard classically valid inference patterns in BAT.

Premises	Conclusion	$\vdash$?
φ	$\neg\neg\varphi$	Yes
$\neg\neg\varphi$	φ	Yes
$\neg(\varphi \wedge \psi)$	$\neg\varphi \vee \neg\psi$	No
$\neg\varphi \vee \neg\psi$	$\neg(\varphi \wedge \psi)$	No
$\neg(\varphi \vee \psi)$	$\neg\varphi \wedge \neg\psi$	Yes
$\neg\varphi \wedge \neg\psi$	$\neg(\varphi \vee \psi)$	Yes
$\varphi \wedge \psi$	$\psi \wedge \varphi$	Yes
$\varphi \vee \psi$	$\psi \vee \varphi$	No
$\varphi \to \psi$	$\neg\psi \to \neg\varphi$	No
$\varphi \wedge \psi$	$\varphi \vee \psi$	Yes
φ	$\psi \to \varphi$	Yes
$\varphi \wedge (\psi \vee \chi)$	$(\varphi \vee \psi) \wedge (\varphi \vee \chi)$	No
$\varphi \vee (\psi \wedge \chi)$	$(\varphi \wedge \psi) \vee (\varphi \wedge \chi)$	No
$(\varphi \vee \psi) \wedge (\varphi \vee \chi)$	$\varphi \wedge (\psi \vee \chi)$	No
$(\varphi \wedge \psi) \vee (\varphi \wedge \chi)$	$\varphi \vee (\psi \wedge \chi)$	No
$\varphi \to \psi, \psi \to \chi$	$\varphi \to \chi$	No
$\neg\psi$	$\neg(\varphi \wedge (\varphi \to \psi))$	No
φ	$\neg[(\varphi \to \psi) \wedge \neg\psi]$	No
$\varphi \vee \psi, \neg\varphi$	ψ	Yes
$\varphi \to \psi, \neg\psi$	$\neg\varphi$	Yes
$\neg\psi \wedge (\varphi \to \psi)$	$\neg\varphi$	Yes
$\varphi \wedge (\varphi \to \psi)$	ψ	Yes
$\varphi, (\varphi \to \psi)$	ψ	Yes
$\varphi \to \psi$	$(\varphi \wedge \lambda) \to \psi$	No
$\varphi \to \psi$	$\varphi \to (\psi \vee \lambda)$	No
$\varphi \wedge (\psi \wedge \chi)$	$(\varphi \wedge \psi) \wedge \chi$	Yes
$\varphi \vee (\psi \vee \chi)$	$(\varphi \vee \psi) \vee \chi$	No
$(\varphi \wedge \psi) \wedge \chi$	$\varphi \wedge (\psi \wedge \chi)$	Yes
$(\varphi \vee \psi) \vee \chi$	$\varphi \vee (\psi \vee \chi)$	No

Notice that *Modus Ponens* works in both formulations, while its contraposed form fails. Similarly, *Modus Tollens* works in both forms, while its contraposed form fails. This entails:

Theorem 18. *It is not generally the case that if* $\varphi \not\vdash \psi$ *then* $\neg\psi \not\vdash \neg\varphi$.

Observe that disjunction is neither commutative nor associative. Take the assignment v where all propositional variables have value n and consider two formulas: $\varphi = p \vee q$ and $\psi = q \vee p$. As far as φ and ψ are concerned, there are four possible ways to extend this assignment:

$$e_v^1(\varphi) = n = e_v^1(\psi)$$
$$e_v^2(\varphi) = 1, e_v^2(\psi) = n$$
$$e_v^3(\varphi) = n, e_v^3(\psi) = 1$$
$$e_v^4(\varphi) = 1 = e_v^4(\psi).$$

BAT is too weak to eliminate extensions (e_v^1, e_v^2, e_v^3), in which φ and ψ obtain different values, and which show that neither $\varphi \not\vdash \psi$, nor $\psi \not\vdash \varphi$. Thus, it needs to be strengthened.

2.5 Strengthening BAT

Usually, to obtain a stronger logic from a logic with a non-deterministic semantics we have to limit the range of available possible extensions of given assignments.[5] We would like to propose our own solution to this problem in terms of either enriching one logic by another one or by an additional closure condition.

Definition 19. Let L be a logic. We say that a BAT-evaluation e belongs to the L-filtered set of BAT-evaluations just in case the following conditions hold:

1. For any two formulas φ, ψ, if $\models_L \varphi \equiv \psi$ then $e(\varphi) = e(\psi)$,
2. For any L-tautology φ, $e(\varphi) = 1$,
3. For any L-countertautology φ, $e(\varphi) = 0$.

We will focus on the case where L is classical logic (L=CL) in the extended language with B,[6] and we simply use $\models$ to denote the classical consequence relation. By $\Gamma \not\vdash_{CL} \varphi$ we will mean that for any evaluation e in the CL-filtered set

[5] The most common way to strengthen a non-deterministic logic is to use the level-evaluation method (Coniglio et al., 2015) Due to simplicity, we prefer our method.

[6] Technically speaking, propositional logic is not defined for the language containing operator B. Yet, it is rather straightforward what we mean. We mean a logic which treats every formula whose main operator is B as a propositional variable. Semantically, one can interpret B by an n-matrix:

B	φ
{1,0}	1
{1,0}	0

of BAT-evaluations if $e(\psi) = 1$ for all $\psi \in \Gamma$ then $e(\varphi) = 1$. The resulting logic is called CLBAT.

We may also be inclined to strengthen BAT in a different manner. A quite intuitive way to go is to close BAT under classical consequence. It can be done by the following condition:

Definition 20 (Closure condition)**.** An extension of BAT (in $\mathcal{L}_{\mathtt{B}}$) satisfies the closure condition just in case for all $\mathcal{L}_{\mathtt{B}}$-formulas $\varphi_1, \varphi_2, \cdots, \varphi_k, \psi$ such that

$$\varphi_1, \varphi_2, \cdots, \varphi_k \models \psi,$$

where $\models$ is the classical consequence relation for $\mathcal{L}_{\mathtt{B}}$, for any BAT-evaluation e_v, if $e_v(\mathtt{B}\varphi_i) = 1$ for any $0 < i \leq k$, then $e_v(\mathtt{B}\psi) = 1$.

The result of closing BAT-logic under the closure condition will be called CABAT and its consequence relation will be denoted by $\vdash_C$.

It turned out that the above conditions are equivalent, and so are the resulting logics:

Theorem 21. $\Gamma \vdash_C \varphi$ *iff* $\Gamma \vdash_{CL} \varphi$.

Proof. We will show that the set of CL-filtered BAT-evaluations respects the closure condition and that the set of evaluations for which the closure condition holds is exactly the set of CL-filtered BAT-evaluations.

$\Rightarrow$: Let $\Gamma = \{\varphi_1 \ldots \varphi_n\}$ and φ be such that $\Gamma \models \varphi$. We have to show that for any CLBAT-evaluation e, if $e(\mathtt{B}\varphi_1) = e(\mathtt{B}\varphi_2) = \ldots = e(\mathtt{B}\varphi_n) = 1$ then $e(\mathtt{B}\varphi) = 1$. Assume the antecedent. By the deduction theorem for classical propositional logic we know that $\models \bigwedge_{i=1}^{i=n} \varphi_i \rightarrow \varphi$. By the assumption and the definition of CLBAT-evaluation, $e(\bigwedge_{i=1}^{i=n} \varphi_i \rightarrow \varphi) = 1$. Since any CLBAT-evaluation which assigns 1 to all conjuncts has to assign 1 to the whole conjunction, we have $e(\bigwedge_{i=1}^{i=n} \varphi_i) = 1$. By the matrix for implication it follows that $e(\varphi) = 1$. By the matrix of $\mathtt{B}$, we have $e(\mathtt{B}\varphi) = 1$.

$\Leftarrow$: To show that any CABAT-evaluation e is also a CLBAT-evaluation, since both sets are subsets of BAT-evaluation, we only need to check that CABAT-evaluations respect the filtration conditions.

We will start with the third condition. Let φ be a formula in $\mathcal{L}_{\mathtt{B}}$. First, suppose that for any classical evaluation e, $e(\varphi) = 0$. It follows that for any classical evaluation $e(\neg\varphi) = 0$, so $\models \neg\varphi$. We have to show that φ has value 0 in any CABAT-evaluation. By the closure condition $\vdash_C \mathtt{B}\neg\varphi$. By the matrix for $\mathtt{B}$, $\vdash_C \neg\varphi$. So any CABAT-evaluation assigns 0 to formula φ.

Next, consider the second condition. Let φ be a formula in $\mathcal{L}_{\mathtt{B}}$. First, suppose that for any classical evaluation e, $e(\varphi) = 1$. We have to show that φ has value 1 in any CABAT-evaluation. By the closure condition $\vdash_C \mathtt{B}\varphi$. By the matrix for $\mathtt{B}$, $\vdash_C \varphi$. So any CABAT-evaluation assigns 1 to formula φ.

Finally, consider the first condition. Suppose now that for any classical evaluation $e(\varphi) = e(\psi)$. In other words, $\models \varphi \equiv \psi$. By the deduction theorem, $\varphi \models \psi$, $\psi \models \varphi$, $\neg\psi \models \neg\varphi$ and $\neg\varphi \models \neg\psi$. We want to show that for any CABAT-evaluation e_c, $e_c(\varphi) = e_c(\psi)$. By the closure condition, $\mathtt{B}\varphi \vdash_C \mathtt{B}\psi$, $\mathtt{B}\psi \vdash_C \mathtt{B}\varphi$,

$\mathtt{B}\neg\psi \not\vdash_C \mathtt{B}\neg\varphi$ and $\mathtt{B}\neg\varphi \not\vdash_C \mathtt{B}\neg\psi$. Thus, by the matrix for B we have $\varphi \not\vdash_C \psi$, $\psi \not\vdash_C \varphi$, $\neg\varphi \not\vdash_C \neg\psi$ and $\neg\psi \not\vdash_C \neg\varphi$.

We will consider three cases: $e_c(\varphi) = 1$, $e_c(\varphi) = 0$ and $e_c(\varphi) = n$. We will start with the first case. It follows from $\varphi \not\vdash_C \psi$ that $e_c(\psi) = 1$, thus by the matrix for B, $e_c(\mathtt{B}\psi) = 1$.

For the second case if $e_c(\varphi) = 0$, then $e_c(\neg\varphi) = 1$, thus $e_c(\neg\psi) = 1$, so $e_c(\psi) = 0$, which implies by the matrix for B that $e_c(\mathtt{B}\psi) = 1$.

In the third case note that if $e_c(\varphi) = n$ then $e_c(\psi) = n$ because otherwise by an analogous argument to the ones above from $e_c(\psi) \neq n$ we would have that $e_c(\varphi)$ is either 1 or 0, which contradicts the assumption. □

Given that both CABAT and its internal logic are closed under classical consequence, all the worries about syntactic sensitivity that applied to BAT disappear.

2.6 Properties of CABAT

Quite trivially, CABAT is strictly stronger than BAT. The first interesting thing to see is that the deduction theorem is not generally valid in CABAT:

Theorem 22. *If* $\not\vdash_C \varphi \to \psi$ *then* $\varphi \not\vdash_C \psi$ *but it is not always the case that* $\varphi \not\vdash_C \psi$ *implies* $\not\vdash_C \varphi \to \psi$.

In CABAT implications are stronger than the corresponding consequence relation, simply because the consequence relation informs us only about those evaluations in which all the premises have value 1. For instance, the consequence relation $\varphi \not\vdash_C \psi$ does not determine the value of the implication $\varphi \to \psi$ when both ψ and φ have value n. On the other hand, $\not\vdash_C \varphi \to \psi$ uniquely determines the value of the implication under the previous assignment.

Lack of the deduction theorem makes the difference when we look at inference patterns with provability operator. Usually, principles for provability are valid in CABAT as consequence relations whereas their implicational formulations may be invalid. We are not terribly worried about that, since given our reading $\varphi \not\vdash_C \psi$ means that if φ is informally provable then ψ is and this is the phrase we intended to formalize. On the other hand, $\not\vdash_C \varphi \to \psi$ says if φ is informally provable, then so is ψ and *if the antecedent is undecidable then the consequent is either provable or independent*, which is a stronger claim than $\varphi \not\vdash_C \psi$.

Now we will take a look at some schemas involving the provability predicate. Intuitively, informal provability commutes with conjunction but not with disjunction. The fact that either φ or ψ is informally provable does not imply that we can prove either the first or the second disjunct. Of course, the consequence relation in the opposite direction ($\mathtt{B}\varphi \vee \mathtt{B}\psi \not\vdash_C \mathtt{B}(\varphi \vee \psi)$) should hold.

Provided our reading of the consequence relation, $\mathtt{B}\varphi \not\vdash_C \varphi$ may be seen as a certain version of the reflection schema, which is definitely a sound principle for informal provability.

The fact that a certain statement is informally provable is itself an evidence for the informal provability of the provability of that statement. Thus, iterative

principles allowing to either add or subtract the $\mathtt{B}$ operator from the beginning of a formula are also natural.

The table below summarizes which inference patterns are valid in CABAT and whether the principle, according to us, is intuitive or not:

Principle	Valid?	Intuitive?
$(\mathtt{B}\varphi \wedge \mathtt{B}\psi) \vdash_C \mathtt{B}(\varphi \wedge \psi)$	Yes	Yes
$\mathtt{B}(\varphi \wedge \psi) \vdash_C (\mathtt{B}\varphi \wedge \mathtt{B}\psi)$	Yes	Yes
$\mathtt{B}(\varphi \vee \psi) \vdash_C (\mathtt{B}\varphi \vee \mathtt{B}\psi)$	No	No
$(\mathtt{B}\varphi \vee \mathtt{B}\psi) \vdash_C \mathtt{B}(\varphi \vee \psi)$	No	?
$\varphi \vdash_C \mathtt{B}\varphi$	Yes	Yes
$\mathtt{B}\varphi \vdash_C \varphi$	Yes	Yes
$\mathtt{B}\varphi \vdash_C \neg\mathtt{B}\neg\varphi$	Yes	Yes
$\mathtt{B}\varphi \vdash_C \mathtt{BB}\varphi$	Yes	Yes
$\mathtt{BB}\varphi \vdash_C \mathtt{B}\varphi$	Yes	Yes
$\mathtt{B}(\varphi \to \psi) \vdash_C (\mathtt{B}\varphi \to \mathtt{B}\psi)$	No	?
$\mathtt{B}(\varphi \to \psi), \mathtt{B}\varphi \vdash_C \mathtt{B}\psi$	Yes	Yes
$\mathtt{B}(\varphi \wedge \neg\varphi) \vdash_C \mathtt{B}(\psi)$	Yes	Yes
$\mathtt{B}\varphi \vee \mathtt{B}\neg\varphi$	No	No
$\mathtt{B}\varphi \vee \neg\mathtt{B}\varphi$	Yes	Yes
$\neg\mathtt{B}\varphi \vdash_C \mathtt{B}(\neg\varphi)$	No	No
$\mathtt{B}(\neg\mathtt{B}\varphi) \vdash_C \mathtt{B}(\neg\varphi)$	No	No
$\mathtt{B}(\neg\mathtt{B}\neg\varphi) \vdash_C \neg\mathtt{B}(\neg\mathtt{B}\varphi)$	No	No

One thing that might seem worrying is that

$$(\mathtt{B}\varphi \vee \mathtt{B}\psi) \not\vdash_C \mathtt{B}(\varphi \vee \psi)$$

After all, if φ is informally provable, shouldn't $\varphi \vee \psi$ also be informally provable? This worry, however, stems from the fact that the provability of a disjunction in CABAT says something weaker than that one of its disjuncts is provable — after all, $\varphi \vee \neg\varphi$ is going to be informally provable without either φ or $\neg\varphi$ being informally provable. So, we submit, the intuition should be rather captured by requiring that the following should hold:

$$\mathtt{B}\varphi \vdash_C \mathtt{B}(\varphi \vee \psi)$$
$$\mathtt{B}\psi \vdash_C \mathtt{B}(\varphi \vee \psi)$$

and indeed, they do.

Another worry might be that the following asymmetry between at least *prima facie* close cousins can be observed:

$$\mathtt{B}(\varphi \to \psi) \not\vdash_C (\mathtt{B}\varphi \to \mathtt{B}\psi) \qquad \text{(Fake K)}$$
$$\mathtt{B}(\varphi \to \psi), \mathtt{B}\varphi \vdash_C \mathtt{B}\psi \qquad \text{(Real K)}$$

The answer is, however, that putting $\mathtt{B}\varphi \to \mathtt{B}\psi$ on the right-hand side of $\vdash_C$ doesn't adequately capture the intuition that *if φ is informally provable, then so is ψ*. For $\mathtt{B}\varphi \to \mathtt{B}\psi$ actually contains more information than that. *If*

φ *is informally provable, then so is* ψ tells us only what happens when φ (and so, $\mathtt{B}\varphi$) is informally provable, while the provability of $\mathtt{B}\varphi \to \mathtt{B}\psi$ puts further constraints on what happens if φ is not informally provable (for instance, that if it is undecidable, ψ cannot be refutable). It's (Real K) and not (Fake K) that properly captures the underlying intuition.

2.7 CABAT and provability

Now, let's see the difference between using $\vdash_C$ and its provability operator on the one hand, and using Peano Arithmetic and its standard provability predicate (or the modal logic of provability GL and its provability operator, for that matter) on the other.

Quite crucially, we may want to see which principles that hold for standard formal provability predicates hold for operator $\mathtt{B}$ as well.

First, recall Hilbert-Bernays conditions for PA:

$$\text{PA} \vdash \varphi \Rightarrow \text{PA} \vdash Bew(\ulcorner\varphi\urcorner) \tag{HB1}$$

$$\text{PA} \vdash Bew(\ulcorner\varphi \to \psi\urcorner) \to (Bew(\ulcorner\varphi\urcorner) \to Bew(\ulcorner\psi\urcorner)) \tag{HB2}$$

$$\text{PA} \vdash Bew(\ulcorner\varphi\urcorner) \to Bew(\ulcorner Bew(\ulcorner\varphi\urcorner)\urcorner) \tag{HB3}$$

In the arithmetical setting standard Hilbert-Bernays conditions allow one to prove Löb's theorem:

Theorem 23. *If* $\mathbf{PA} \vdash Bew(\ulcorner\varphi\urcorner) \to \varphi$ *then* $\mathbf{PA} \vdash \varphi$.

Since we'll want to make a point about how the standard proofs of the theorems that we'll discuss proceed, we'll go over them quickly.

Proof. Suppose $\mathbf{PA} \vdash Bew(\ulcorner\varphi\urcorner) \to \varphi$. By the diagonal lemma there is a formula such that $\mathbf{PA} \vdash \lambda \equiv (Bew(\ulcorner\lambda\urcorner) \to \varphi)$. Now, arguing inside Peano Arithmetic we get:

$$\lambda \to (Bew(\ulcorner\lambda\urcorner) \to \varphi) \tag{2.1}$$

$$Bew(\ulcorner\lambda \to (Bew(\ulcorner\lambda\urcorner) \to \varphi)\urcorner) \tag{2.2}$$

$$Bew(\ulcorner\lambda\urcorner) \to (Bew(\ulcorner Bew(\ulcorner\lambda\urcorner)\urcorner) \to Bew(\ulcorner\varphi\urcorner)) \tag{2.3}$$

$$Bew(\ulcorner\lambda\urcorner) \to Bew(\ulcorner Bew(\ulcorner\lambda\urcorner)\urcorner) \tag{2.4}$$

$$Bew(\ulcorner\lambda\urcorner) \to Bew(\ulcorner\varphi\urcorner) \tag{2.5}$$

$$Bew(\ulcorner\lambda\urcorner) \to \varphi \tag{2.6}$$

$$\lambda \tag{2.7}$$

$$Bew(\ulcorner\lambda\urcorner) \tag{2.8}$$

$$\varphi \tag{2.9}$$

□

(2.2) is obtained by neccesitation. Next, we use the second Hilbert-Bernay's condition to distribute provability over implication twice to obtain (2.3). Line (2.4) is the third Hilbert-Bernays condition thanks to which we obtain (2.5) from

(2.3). By the assumption that $\mathbf{PA} \vdash Bew(\ulcorner\varphi\urcorner) \to \varphi$ we obtain (2.6). Applying detachment to the sentence generated by the diagonal lemma we get (2.7). Then, by necessitation and *modus ponens*, we obtain the last two lines.

Löb's theorem is not an intuitively sound principle for informal provability. There is no reason to suppose that only those instances of the reflection schema hold for which φ is already a theorem. The intuitions are rather clear that all instances of reflection are plausible.

In the arithmetical setting we get Löb's theorem as a side-effect of the diagonal lemma. It is not something that we would like to postulate as an interesting and independently motivated principle. Rather, it is an unwanted surprising consequence. It is also one of the reasons why we cannot consistently put all the instances of the reflection schema together with HB conditions in the classical setting.

In CABAT we have certain versions of HB conditions:

$$\varphi \mathrel{\text{✒}}_C \mathtt{B}\varphi \tag{HB1$'$}$$

$$\mathtt{B}(\varphi \to \psi), \mathtt{B}\varphi \mathrel{\text{✒}}_C \mathtt{B}\psi \tag{HB2$'$}$$

$$\mathtt{B}\varphi \mathrel{\text{✒}}_C \mathtt{BB}\varphi \tag{HB3$'$}$$

The first condition in CABAT is a bit stronger, since it is not restricted only to theorems. The condition starts to be intuitive as soon as you recall the interpretation of $\varphi \mathrel{\text{✒}}_C \psi$, which says that if φ is informally provable then so is ψ. Some may be worried that the above formulation of HB1, in some sense, allows to go from premises which are true (and may not be theorems) to premises which are theorems. But as we explained, according to our reading formulas on the left hand side of $\mathrel{\text{✒}}_C$ are not true but informally provable. So the principle allows only to go from informally provable premises to informally provable premises having informal provability expressed in the object language.

One interesting question is whether the above conditions are enough to prove Löb's theorem. The key observation, in the standard proof, is that once the premises, including the one produced by an application of the diagonal lemma, are listed, the theorem follows by classical propositional logic. So it seems that the issue can be handled at the propositional level.

The natural way to go about the translation is this. We translate both *Bew* and $\vdash$ as $\mathtt{B}$. It is a standard practice to translate them using a single symbol (see Boolos, 1993).

Slightly more challenging is the question how to translate implications from the language of **PA**. The straightforward approach is to translate them as material implications in $\mathcal{L}_{\mathtt{B}}$.

But we think it will not do justice to the original theorem. The deduction theorem does not hold for CABAT. Implications are stronger claims than consequence claims and are much harder to prove. Thus, whenever possible, we will translate $\varphi \to \psi$ in the conclusions as $\varphi \mathrel{\text{✒}}_C \psi$. We leave implications in the premises, especially within the scope of $\mathtt{B}$. But this is not a cheap way for us to avoid an undesired consequence: by leaving material implications in the premises we make them as strong as we can.

As for sentences produced by the application of Diagonal Lemma we will build them to the assumptions.

Fact 24 (Löb's theorem failure)**.**

$$\mathtt{B}(\mathtt{B}\varphi \to \varphi), \mathtt{B}(\lambda \to (\mathtt{B}\lambda \to \varphi)), \mathtt{B}((\mathtt{B}\lambda \to \varphi) \to \lambda) \not\vdash_C \mathtt{B}\varphi.$$

Proof. Just take an assignment $v(\varphi) = v(\lambda) = n$ and extend it to an evaluation where for each implication if it is possible to choose n, it should be chosen. It is easily verifiable that all the premises have value 1, and yet the conclusion has value n, while all the constraints on valuations remain satisfied. □

Why doesn't the standard argument work? Suppose e_v gives value 1 to all the premises. From $e_v(\mathtt{B}(\lambda \to (\mathtt{B}\lambda \to \varphi))) = 1$, we have $e_v(\mathtt{B}\lambda \to (\mathtt{BB}\lambda \to \mathtt{B}\varphi)) = 1$. Now, in the standard proof we use the fact that $e_v(\mathtt{B}\varphi \to \mathtt{BB}\varphi) = 1$, but we cannot do that here, since in general the previous formula is not a CABAT-tautology.

In other words, it is not the case that only those instances of the reflection schema are provable for which φ is already a theorem. The lack of Löb's theorem is rather promising since it leaves open the possibility for adding all the instances of the reflection schema consistently.

We will take a quick look at two other theorems related to provability and the reflection schema.

As we already stated, there is a problem with the reflection schema in the standard setting. It is impossible to add all instances of the schema and at the same time have all Hilbert-Bernays conditions. This is shown by the Montague's paradox:

Theorem 25. *Peano Arithmetic, if consistent, cannot contain (or be consistently extended to contain) a (possibly complex) predicate for which all Hilbert-Bernays conditions and all instances of the reflection schema hold.*

Proof. Suppose that there is such a predicate, call it P. We use a natural deduction system. Argue inside the theory:

1. $\lambda \equiv P(\ulcorner\neg\lambda\urcorner)$	Diagonal lemma
1.1 λ	Hypothesis
1.2 $P(\ulcorner\neg\lambda\urcorner)$	equivalence elimination: 1,1.1
1.3 $\neg\lambda$	modus ponens and reflection schema: 1.2
2. $\neg\lambda$	reductio ad absurdum: 1.1 $\to$ 1.3
3. $P(\ulcorner\neg\lambda\urcorner)$	HB 1
4. $\neg P(\ulcorner\neg\lambda\urcorner)$	1, 2
5. contradiction	3, 4

□

To rephrase the above theorem, it is impossible, given Hilbert-Bernays conditions, to extend the theory with the inverse of the implication corresponding to the necessitation rule. The same goes for the implication directly corresponding to the necessitation rule:

Theorem 26. *Peano Arithmetic, if consistent, cannot contain (or be consistently extended to contain) a (possibly complex) predicate for which all Hilbert-Bernays conditions and all instances of $\varphi \to P(\ulcorner\varphi\urcorner)$ (Provabilitation) hold and is closed under the co-necessitation rule: if $P(\ulcorner\varphi\urcorner)$ then φ.*

Proof. Suppose that there is such a predicate, call it P. We use a natural deduction system. Argue inside the theory:

1. $\neg P(\ulcorner\kappa\urcorner) \equiv \kappa$	the diagonal lemma
1.1 κ	conditional assumption
1.2 $\neg P(\ulcorner\kappa\urcorner)$	equiv elimination: 1, 1.1
1.3 $\kappa \to P(\ulcorner\kappa\urcorner)$	instance of provabilitation for κ
1.4 $\neg\kappa$	MTT: 1.3, 1.2
2. $\neg\kappa$	conditional assumption discharge: 1.1 $\leftarrow$ 1.4
3. $P(\ulcorner\kappa\urcorner)$	equiv elimination: 1,2
4 κ	co-necessitation: 3
5. contradiction	2,4

□

The moral is that the price for all Hilbert-Bernays conditions together with all instances of the reflection schema on the one hand, or all the instances of provabilitation one the other (assuming co-necessitation) is too high in the standard setting.

Fact 27 (Montague)**.** $\mathtt{B}(\mathtt{B}\lambda \to \lambda), \mathtt{B}(\mathtt{B}\neg\lambda \to \neg\lambda), \mathtt{B}(\mathtt{B}\neg\lambda \equiv \lambda) \not\vdash_C \lambda \wedge \neg\lambda$.

Proof. We can omit all occurrences of $\mathtt{B}$, in all formulas of the form $\mathtt{B}\varphi$. Thus, we can omit the outmost left $\mathtt{B}$ in all the premises. Let e_v be an evaluation under all the premises have value 1. Then, since $e_v(\mathtt{B}\neg\lambda \to \neg\lambda) = 1 = e_v(\mathtt{B}\neg\lambda \to \lambda)$, it follows that $e_v(\neg\mathtt{B}\neg\lambda) = 1$. From $\mathtt{B}\neg\lambda \equiv \lambda$ and $e_v(\neg\mathtt{B}\neg\lambda) = 1$ by classical logic we have $e_v(\neg\lambda) = 1$ and $e_v(\mathtt{B}\neg\lambda) = 0$. Which is already a contradiction, since $e_v(\neg\lambda) = 1$ implies that $e_v(\mathtt{B}\neg\lambda) = 1$.

□

Fact 28 (Dual Montague)**.** *The following consequence still holds:* $\mathtt{B}(\lambda \to \neg\mathtt{B}\lambda), \mathtt{B}(\neg\mathtt{B}(\lambda) \to \lambda), \mathtt{B}(\lambda \to \mathtt{B}\lambda), \mathtt{B}(\neg\lambda \to \mathtt{B}\neg\lambda) \not\vdash_C \lambda \wedge \neg\lambda$.

Proof. Similarly we can omit all occurrences of $\mathtt{B}$, in all formulas of the form $\mathtt{B}\varphi$. Let e_v be an evaluation under which all the premises have value 1. Then, since $e_v(\lambda \to \mathtt{B}\lambda) = 1 = e_v(\lambda \to \neg\mathtt{B}\lambda)$, it follows that $e_v(\neg\lambda) = 1$. Observe that $e_v(\neg\mathtt{B}\lambda \to \lambda)$ implies, by *modus tollens*, $e_v(\neg\neg\mathtt{B}\lambda) = 1$ finally resulting in $e_v(\mathtt{B}\lambda) = 1 = e_v(\lambda)$. Contradiction. □

The moral of the above is that, even in CABAT there is no possibility of having full reflection schema. Yet, we have more reflection than usually:

Fact 29. *Suppose either* $\not\vdash_C \varphi$ *or* $\not\vdash_C \neg\varphi$*. It follows that* $\not\vdash_C \mathtt{B}\varphi \to \varphi$.

Proof. Take any evaluation e_v, it is clear that since $e_v(\varphi) = 1$ or $e_v(\varphi) = 0$, we have $e_v(\neg\mathtt{B}\varphi \vee \varphi) = 1$, which shows that $e_v(\mathtt{B}\varphi \to \varphi) = 1$. □

This is also a signal that our initial intuition behind CABAT is not completely insane: with CABAT in the background it is possible to have more of the reflection schema, but not all the instances of provabilitation (which is less intuitive for informal provability).

Even if we add reflection for both λ and $\neg\lambda$, where λ states the provability of its own negation, CABAT proves that λ is informally undecidable.

Fact 30 (Reflection and provability)**.** *The following consequence holds:* $\mathtt{B}(\mathtt{B}\lambda \to \lambda), \mathtt{B}(\mathtt{B}\neg\lambda \to \neg\lambda), \mathtt{B}(\mathtt{B}(\neg\lambda) \to \lambda), \mathtt{B}(\lambda \to \mathtt{B}(\neg\lambda)) \not\vdash_C \neg\mathtt{B}\neg\lambda \wedge \neg\mathtt{B}\lambda$.

Proof. As usual we omit the left-hand side B. Take any evaluation for which $e_v(\varphi) = 1$ for all φ which are premises. The instance of the reflection schema $e_v(\mathtt{B}\neg\lambda \to \neg\lambda)$ combined with $\mathtt{B}\neg\lambda \to \lambda$ gives $e_v(\neg\mathtt{B}\neg\lambda) = 1$. This implies $e_v(\neg\lambda) = 1$, thus $e_v(\neg\mathtt{B}\lambda) = 1$. In other words every evaluation e_v which gives 1 to all the premises also gives 1 to $\neg\mathtt{B}\neg\lambda \wedge \neg\mathtt{B}\lambda$. □

That's for a sentence which says that its own negation is informally provable. Another interesting pet worth playing with is what we'll call *informal Gödel sentence*: a sentence which says of itself that it is not *informally* provable. The first thing to observe is that its formalization in CABAT doesn't lead to contradiction:

Fact 31 (Informal Gödel sentence)**.** $\gamma \to \neg B\gamma, \neg B\gamma \to \gamma \not\vdash_C \gamma \wedge \neg\gamma$

Proof. Because of the closure condition, $\not\vdash_C B(\neg(\gamma \wedge \neg\gamma))$, and so $\not\vdash_C \neg(\gamma \wedge \neg\gamma)$ and $e_v(\gamma \wedge \neg\gamma)$ has to be 0, independently of what $e_v(\gamma)$ is, in particular it is possible that $e_v(\gamma) = n$. So now the only thing that we need to check is whether it's possible that all the premises have value 1. Indeed, if $e_v(\gamma) = n$, there is no problem with assuming that e_v assigns 1 to both premises. After all, they are just implications whose at least one argument has value n, and such implications can be assigned value 1 (see the matrix for $\to$). □

However, as soon as we add either provabilitation or reflection contradiction follows:

Fact 32 (Informal Gödel with provabilitation)**.** *The following holds:*

$$\gamma \to \mathtt{B}\gamma, \gamma \to \neg B\gamma, \neg B\gamma \to \gamma \not\vdash_C \gamma \wedge \neg\gamma.$$

Proof. Again, the proof proceeds by showing that no evaluation can assing 1 to all the premises. For contradiction, suppose e_v is an evaluation which assigns 1 to all the premises. By the fact that $e_v(\gamma \to \mathtt{B}\gamma) = e_v(\gamma \to \neg B\gamma) = 1$, we have $e_v(\gamma) = 0$. Then, $e_v(\neg\gamma) = 1$. Apply *modus tollens* to the third premise, we have $e_v(\neg B\gamma) = 0$ and $e_v(B\gamma) = e_v(\gamma) = 1$, which is a contradiction. □

The above fact isn't too worrying, because provabilitation doesn't seem too plausible for informal provability to start with. Here's a more interesting case.

Fact 33 (Informal Gödel with reflection)**.** *The following holds in CABAT:*

$$\mathtt{B}\gamma \to \gamma, \mathtt{B}\neg\gamma \to \neg\gamma, \mathtt{B}\gamma \to \neg\gamma, \neg\gamma \to \mathtt{B}\gamma \not\vdash_C \gamma \wedge \neg\gamma.$$

Proof. By the argument used in the previous proof, $\gamma \wedge \neg\gamma$ will always have value 0. So the only way of showing that the consequence holds is to prove that the premises cannot all have value one. For contradiction, assume e_v is an evaluation which assigns 1 to all the premises. Then, by the transitivity of implication in CABAT (guaranteed by the closure condition) we have $e_v(\neg\gamma \to \gamma) = 1$ By the closure condition, $e_v(\neg\gamma \to \gamma) = e_v(\neg\neg\gamma \vee \gamma) = e_v(\gamma \vee \gamma) = e_v(\gamma) = 1$. Thus $e_v(\mathtt{B}\gamma) = 1$ and by *modus ponens* applied to the third premise $e_v(\neg\gamma) = 1$, which is a contradiction. □

One might be worried that the last fact gives rise to a paradox in the vein of (Priest, 2006; Beall, 1999), where the argument is put forward to the effect that informal mathematics is inconsistent, because one can take the sentence:

(γ) γ is not informally provable.

and both the assumption that γ is informally provable, and that it isn't informally provable lead to contradiction.

Paradoxical arguments in natural language aside, notice that given that CABAT consequence relation is defined in terms of informal provability preservation, Fact 33 is to be read: *if the formulae on the left-hand side of* $\vdash_C$ *are informally provable, then so is the formula on the right-hand side.* So, assuming reflection is informally provable, for the paradox to arise we actually have to assume not only that the following is true:

$$\gamma \equiv \neg B\gamma, \tag{γ'}$$

but also that it is *provable in informal mathematics.* This, however, is a stronger assumption that standard paradoxical arguments failed to establish: whether writing down (γ) constitutes an *informal mathematical proof* of (γ') is far from obvious and deserves a separate discussion. These and related fascinating issues, however, lie beyond the scope of this paper.

2.8 Conclusions

Once we intuitively divide mathematical claims into provable, refutable and independent, the question arises as to how these three classes interact with Boolean connectives. This interaction is not straightforward, because facts about whether certain claims are provable, refutable, or independent do not unambiguously determine the status of their Boolean combinations.

This obstacle, however, is not fatal. Once we move to non-deterministic semantics, the basic constraints on how provability, refutability and independence behave with respect to Boolean connectives can be explicated by a formal system: BAT. The constraints captured by BAT matrices are a bit too basic, though. They don't give justice to the fact that informal mathematical provability is closed under classical consequence. Adding this requirement to BAT results in a stronger system, CABAT, which is studied in the remainder of the paper.

CABAT, in contrast with BAT, doesn't fall prey to syntactic sensitivity. CABAT also validates many intuitively plausible and invalidates many intuitively implausible inference patterns for informal provability. Among the invalidated ones, we have Löb's theorem, which when applied to informal provability seems to be making the unintuitive claim that reflection holds only for those statements which are already informally provable. The failure of Löb's theorem makes all the instances of the reflection schema consistent with CABAT.

Chapter 3

Proof systems for BAT consequence relations[1]

Abstract

An ongoing debate about the differences between formal provability in an axiomatic system and informal provability of mathematical claims in mathematics as a whole resulted in the construction of various logics whose main purpose is to capture the inferential behavior of the notion of informal provability, just as multiple logics of formal provability capture the behavior of the concept of formal provability. Known logics of informal provability, based on classical logic, are unable to incorporate all intuitive principles of informal provability (most notably, reflection, which says that whatever is provable is true). One solution to this problem is to treat informal provability as an operator (Shapiro, 1985; Reinhardt, 1986; Koellner, 2016). Another solution is to weaken some of the intuitively adequate principles (Horsten, 2002). Recently, in a yet another approach to the issue, two three-valued non-deterministic logics of informal provability have been developed (Pawlowski and Urbaniak, 2017) to overcome this difficulty. Alas, the logics have been characterized semantically and no proof systems for them are available. The purpose of this paper is to define tree-like proof systems for those logics and to prove the corresponding soundness and completeness theorems.

Keywords. informal provability, non-deterministic logic, proof systems for non-deterministic logics, logic of informal provability, non-deterministic matrices

Acknowledgments.
Research on this paper has been funded by the Research Foundation Flanders (FWO). The author would like to express his gratitude to all those who commented on the earlier versions of this paper, especially to Rafal Urbaniak, Frederik Van De Putte and anonymous referees.

[1] I am the single author and the paper is accepted and forthcoming in the *Logic Journal of the IGPL*.

3.1 Motivations

The main aim of this paper is to provide sound and complete proof systems for two modal logics of provability BAT and CABAT developed and discussed in (Pawlowski and Urbaniak, 2017). We will develop tree-like proof systems in the spirit of (Beth, 1955; Carnielli, 1987; Priest, 2006), and prove soundness and strong completeness.

These logics were developed to model an informal notion of provability in classical mathematics. Roughly speaking, a mathematical claim is informally provable if and only if it can be proved using widely accepted mathematical techniques.[2] It may be the case that commonly accepted methods may change through time, but at each point in time there seems to be a certain core set of widely accepted mathematical techniques which give rise to informal provability.

On the other hand, we have *formal proofs* given by means of a proof system of a fully formalized theory. Proofs in this sense are always relative to a given proof system and a given theory. According to the proponents of *the standard view*[3] (see (Antonutti Marfori, 2010) and (Rav, 1999) for an elaborate discussion on this), the relation between formal and informal concepts of provability is straightforward. Informal proofs are incomplete sketches of formal proofs. In principle, they claim, any informal proof can be converted into a proper proof in a relevant axiomatic system — if one thinks there is one system in which all formal proofs can be obtained, say ZFC, one should say *the* relevant axiomatic system; but this hinges on the particularities of the variant of the view.

Some philosophers argue that the above picture is too simplistic and that the relation between formal and informal proofs is different. Antonutti Marfori (2010) claims that there is no clear algorithm for converting a given informal proof into a proper proof in a relevant axiomatic system. Tanswell (2015) claims that it is not obvious how we can identify different informal proofs with their translations. Rav (1999, 2007) discusses the epistemological and explanatory superiority of informal proofs over formal ones, arguing that it is not convincingly explained by the proponents of the standard view.

Leitgeb (2009) observed that these concepts of proofs are different. While in formal proofs the language is precisely defined and divided according to logical order, informal proofs are stated in a natural language expanded with additional mathematical vocabulary. Moreover, the connection between steps in an informal

[2] I do not claim that mathematics is a unified discipline. In some branches of mathematics some additional mathematical techniques are available whereas these additional techniques may lead to incorrect results in the other branches. Nonetheless, it seems that there is a common core of mathematical ways of proving things which is accepted thorough all its sub-disciplines.

[3] This view is usually shared by mathematicians, for instance Enderton (1977, 10-11)

> It is sometimes said that "mathematics can be embedded in set theory." This means that mathematical objects (such as numbers and differentiable functions) can be defined to be certain sets. And the theorems of mathematics (such as the fundamental theorem of calculus) then can be viewed as statements about sets. Furthermore, these theorems will be provable from our axioms. Hence our axioms provide a sufficient collection of assumptions for the development of the whole of mathematics — a remarkable fact. (In Chapter 5 we will consider further the procedure for embedding mathematics in set theory.)

Also, for a bit more sophisticated version of the standard view, see (Sjögren, 2010).

proof has a different nature than in the formal one. The former often employs steps that are supposed to be intuitively seen as truth-preserving, without explicitly following syntactically formulated rules of inference and the latter is based purely on syntactical proof forming rules.

For this paper the most important difference between formal and informal provability lies in the set of general principles sound for both kinds of provability. An important inference pattern for informal provability is *the reflection schema*.[4] It roughly says that whatever is provable, is true. It is a well-known fact that there is no consistent formal theory extending Peano arithmetic in which all instances of the reflection schema for its own formal provability predicate are provable (see notably (Myhill, 1960; Montague, 1963)). So, it seems that the informal notion of provability cannot be formally represented in the standard setting.

A couple of interesting approaches to this problem have been developed. Shapiro (1985) constructed a theory called *Epistemic Arithmetic* (EA) where informal provability is formalized as an operator not as a predicate. On this approach informal provability is governed by a modal logic S4.[5] Shapiro defined a theorem-preserving translation V from the language of arithmetic based on intuitionistic logic (Heyting arithmetic) to the language of EA by the following:

1. For atomic formulas: $V(\overline{\varphi}) = \Box\varphi$
2. $V(\overline{\varphi \wedge \psi}) = \Box(V(\overline{\varphi})) \wedge \Box(V(\overline{\psi}))$
3. $V(\overline{\varphi \vee \psi}) = \Box(V(\overline{\varphi})) \vee \Box(V(\overline{\psi}))$
4. $V(\overline{\varphi \to \psi}) = \Box(\Box(V(\overline{\varphi})) \to \Box(V(\overline{\psi})))$
5. $V(\overline{\varphi \equiv \psi}) = \Box(\Box(V(\overline{\varphi})) \equiv \Box(V(\overline{\psi})))$
6. $V(\overline{\neg\varphi}) = \Box(\neg\Box(V(\overline{\varphi})))$
7. $V(\overline{\forall x\, \varphi(x)}) = \Box(\forall x\, V(\overline{\varphi(x)}))$
8. $V(\overline{\exists x\, \varphi(x)}) = \Box(\exists x\, V(\overline{\varphi(x)}))$

where $\overline{\varphi}$ means that φ is an intuitionistic formula.[6]

Goodman (1984) proved that the translation V is faithful. His proof was notably simplified by Flagg and Friedman (1986). This theory was further developed in three directions: by considering additional principles such as the

[4] This schema was thoroughly studied in (Beklemishev, 1997, 2003; Arai, 1998).

[5] Recall $S4$ is axiomatized by:

1. $\vdash \Box\varphi \to \varphi$
2. $\vdash \Box\varphi \to \Box\Box\varphi$
3. $\vdash \Box\varphi \equiv \Box\Box\varphi$
4. If $\vdash \varphi$ then $\vdash \Box\varphi$
5. $\vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$.

[6] The overline is important because the meaning of functors in intuitionistic logic is different than in classical.

Epistemic Church thesis (see (Flagg and Friedman, 1986; Halbach and Horsten, 2000)), by extending the language with a truth predicate Stern (2015); Koellner (2016), and by a deeper analysis of the informal provability operator (Horsten, 1994, 1997; Heylen, 2013; Rin and Walsh, 2016).

The collateral damage of this approach is a serious limitation in expressive power. It is no longer possible to quantify over formulas by means of coding. Moreover, the internal logic of the operator in this theories is quite weak because of the existence of this translation.

A somewhat different approach to informal provability was proposed by Horsten (1997). The idea is simple: informal provability remains as a predicate but the set of its intuitive principles is weakened. We split the set of intuitive principles into two, we add some of them to the first arithmetical theory called the basis. Next, we add the rest of the principles to the main theory and we also add all the instances of the bridging schema saying that if something is provable in the basis, it is informally provable in the main theory.

This approach seems to be more promising but it has its own problems. First, there is no principled and independently motivated story explaining where the set of intuitive principles should be split. Morover, Stern (2015) proved that a lot of similar systems are inconsistent. Thus, the approach seems not to be as promising as initially suspected.

Recently, Pawlowski and Urbaniak (2017) proposed an alternative way to build a theory of informal provability.[7] The authors developed non-deterministic three-valued logics which can be used for building a formal theory of informal provability. On this approach the move to predicate level seems viable.[8] The translations of paradoxical theorems blocking the move to predicate level for theories having classical logic in the background do not hold. The systems developed in (Pawlowski and Urbaniak, 2017) have been presented without proof theory, and the goal of this paper is to provide proof systems for these logics.

3.2 BAT and CABAT consequence relations

Let $\mathcal{L}$ be a modal propositional language with $\neg, \wedge, \vee, \rightarrow$ as Boolean connectives and a unary modal operator $\mathtt{B}$ whose intended interpretation is *informally provable*. We will use lower case Latin letters $p, q, r, s, \ldots$ as propositional variables. The definitions of atoms and well-formed formulas are standard. Lower case Greek letters $\varphi, \psi, \chi, \ldots$ are meta-variables for (possibly complex) formulas. We will use Γ (possibly with indices) as a variable for finite sets of propositional WFFs. We will treat $\neg, \vee, B$, as primitive connectives, since the rest of them can be defined[9] in a standard way.

BAT has three values: $0, n, 1$. The intended interpretation of 0 is *informally refutable*, 1 stands for *informally provable* and n stands for *neither*.

[7] This paper is included in this thesis as Chapter 2.

[8] Note that in these logics informal provability is treated as an operator not as a predicate. It's possible to develop a first order versions of these logics where informal provability is a predicate. This lies beyond the scope of this paper.

[9] In the sense of having the same matrices.

A BAT-assignment is a function which assigns values to all propositional variables. To obtain the possible values of a complex formula we use the following matrices:

$\neg$	φ
0	1
n	n
1	0

$\vee$	0	n	1
0	0	n	1
n	n	n/1	1
1	1	1	1

B	φ
1	1
n/0	n
0	0

By x/y we mean that the value of a complex formula is not determined by an assignment and it can be either x or y. BAT-assignments do not uniquely determine values for all complex formulas. A BAT-evaluation is a function which extends an assignment, assigns values to all formulas of $\mathcal{L}$, according to the tables given above. By $\Gamma \vdash\!\!\!\!\sim \varphi$, we mean that any BAT-evaluation which assigns 1 to all formulas in Γ assigns 1 to formula φ.

The rest of the standard Boolean connectives have the following matrices:

$\rightarrow$	0	n	1
0	1	1	1
n	n	n/1	1
1	0	n	1

$\equiv$	0	n	1
0	1	n	0
n	n	0/n/1	n
1	0	n	1

$\wedge$	0	n	1
0	0	0	0
n	0	0/n	n
1	0	n	1

The semantics is motivated by a natural division of mathematical claims into provable, refutable and undecidable. Consider a disjunction of two formulas p, q. If at least one of them is informally provable, then intuitively so is the whole disjunction. We can refute the disjunction only if we can refute both p and q. But from the undecidability of p and q we cannot determinately infer the status of their disjunction. For instance, in mathematical practice the status of the Continuum Hypothesis and its negation is regarded as undecided.[10] But there is an agreement about the status of their disjunction — it is provable, simply because it is a substitution of excluded middle. For some other undecided sentences it is the case that their disjunction is also undecided.[11] For instance, consider the disjunction of the consistency statement of ZFC and the Continuum hypothesis. Similar arguments can be developed for other connectives. Hence, indeterminism is both needed and independently motivated.

I do not want to decide whether there are sentences which are absolutely undecidable — merely to allow for such a possibility. It seems that mathematicians behave as if some mathematical claim are undecidable or independent in a certain sense. For instance, some claims can seem independent because they are

[10] Philosophically speaking the question whether the Continuum Hypothesis is really undecidable is a bit more complex. Simply stating that CH and its negation are undecided does not do justice to the range of contemporary opinions on the topic. Arguments have been proposed in favor of the truth of both of these statements. However, the independence of CH and its negation from both the axioms of ZFC set theory, and its extensions with large cardinal axioms makes them reasonable candidates for undecidable statements.

[11] Note that saying that the disjunction of CH and its negation is informally provable in virtue of being a substitution of excluded middle is something different that claiming that every sentence is either informally provable or refutable. We agree with the former not with the latter.

not provable from the commonly accepted mathematical axioms (the Continuum Hypothesis) or that currently proofs of them are not known. Thus, this informal division of mathematical claims seems to be justified in mathematical practice.

BAT is too weak to be used as a logic of informal provability. For instance, it does not prove that disjunction is symmetric. A rather natural strengthening of BAT is obtained by closing its inner logic under classical logic:

Closure condition:

> For every $\mathcal{L}$-formulas $\varphi_1, \varphi_2, \cdots, \varphi_k, \psi$ such that
>
> $$\varphi_1, \varphi_2, \cdots, \varphi_k \models \psi,$$
>
> where $\models$ is the classical consequence relation in the language with a modal operator, for any BAT evaluation e, if $e(\mathtt{B}\varphi_i) = 1$ for all $0 < i \leq k$, then $e(\mathtt{B}\psi) = 1$.

This closure condition does justice to the notion of informal provability, since doing real proofs we do not question classically correct inferences.

By a CABAT evaluation we will mean any BAT evaluation which respects the above closure condition. We will use $\Gamma \vdash_C \varphi$ to denote CABAT consequence relation.

3.3 Informal provability and Löb's theorem

One of the technical features that distinguish formal and informal provability is the validity of the *reflection schema*. Roughly speaking, it says that if something is provable, then it is true.

In mathematical practice it seems that this principle is presupposed. Usually, the existence of an informal proof (where the connection between two steps may be truth-preservation) is sufficient for taking the claim to be true.

Things get interesting when we look at formal provability in a sufficiently strong arithmetical theory. Let T be a recursively-axiomatizable theory extending Robinson arithmetic. Let Bew_T denote its standard provability predicate. It is well known that we have the following Hilbert-Bernays derivability conditions for it:

$$T \vdash \varphi \Rightarrow T \vdash Bew_T(\ulcorner \varphi \urcorner) \tag{HB1}$$

$$T \vdash Bew_T(\ulcorner \varphi \to \psi \urcorner) \to (Bew_T(\ulcorner \varphi \urcorner) \to Bew_T(\ulcorner \psi \urcorner)) \tag{HB2}$$

$$T \vdash Bew_T(\ulcorner \varphi \urcorner) \to Bew_T(\ulcorner Bew_T(\ulcorner \varphi \urcorner) \urcorner) \tag{HB3}$$

The above conditions are sufficient for proving Löb's theorem. Suppose that there is an arithmetical formula θ which behaves as if it were a provability predicate (it satisfied the above HB conditions), then we can prove the following:

Theorem 34. *If* $T \vdash \theta(\ulcorner \varphi \urcorner) \to \varphi$ *then* $T \vdash \varphi$.

The obvious consequence of the above is the fact that it is impossible to have all HB conditions together with the reflection schema. Löb's theorem informs us

how much reflection is allowed for formal provability: we can only have reflection for theorems.

There is little to no discussion about the philosophical significance of Löb's theorem. It seems that it is commonly accepted in virtue of provability obeying HB conditions. Yet, its independent philosophical motivation seems fishy. It seems that there is no reason to accept only those instances of the reflection schema for which the code of a formula used in the schema is a code of a theorem. This restriction is completely artificial and stems from the fact that a particular arithmetical theory is capable of proving Löb's theorem.

One way to go about this problem is to drop Löb's theorem and weaken some of the HB conditions and add more reflection. This is exactly the direction that was taken during the formulation of CABAT. Consider the following properties of CABAT:

Fact 35. *The following hold for CABAT:*

1. $\mathtt{B}(\mathtt{B}\phi \to \phi), \mathtt{B}(\lambda \equiv (\mathtt{B}\lambda \to \phi)) \not\vdash_C \mathtt{B}\phi$.
2. $\mathtt{B}(\mathtt{B}\lambda \to \lambda), \mathtt{B}(\mathtt{B}\neg\lambda \to \neg\lambda), \mathtt{B}(\mathtt{B}(\neg\lambda) \equiv \lambda) \vdash_C \lambda \wedge \neg\lambda$.
3. $\mathtt{B}(\lambda \equiv \neg\mathtt{B}\lambda), \mathtt{B}(\lambda \to \mathtt{B}\lambda), \mathtt{B}(\neg\lambda \to \mathtt{B}\neg\lambda) \vdash_C \lambda \wedge \neg\lambda$.
4. *If* $\vdash_C \varphi$ *or* $\vdash_C \neg\phi$ *then* $\vdash_C \mathtt{B}\varphi \to \varphi$,
5. $\varphi \vdash_C \mathtt{B}\varphi$ *and* $\mathtt{B}\varphi \vdash_C \varphi$

1 shows that a translation of Löb's theorem does not hold in CABAT.[12] The first premise is a translation of the antecedent of Löb's theorem ($\mathbf{T} \vdash \mathtt{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner) \to \varphi$) and the remaining premise is a translation of the application of the diagonal lemma to a formula $\mathtt{Pr}_{\mathbf{T}}(x) \to \varphi$. 2 says that it is not possible to extend CABAT with a translation of the *reflection schema*. The first two premises are translations of the reflection schema and the last one is a translation of the application of the diagonal lemma. Similarly, the third thing on the list states that one cannot consistently add all the instances of the provabilitation schema (if something is true, it is provable). Again, the first two premises are translations of the provabilitation schema for $\lambda, \neg\lambda$ and the last one a translation of the application of the diagonal lemma. Item 4 shows that the amount of reflection by default available in CABAT is greater than for the standard provability predicate. The reflection schema is not only restricted for theorems. The last item on the list shows that CABAT consequence relation preserves provability both ways (meaning a stronger version of NEC and CONEC are valid).[13] This is a good sign. Initially what we wanted for informal provability is the reflection schema which we have in a certain form.

[12] To be fair if we add to the premise set a formula $\mathtt{B}\neg\lambda \to \neg\lambda$ then the consequence does hold. But we are not interested here in an arbitraty instance of the reflection schema but in the translation of the standard proof of Löb's theorem. Thus, in general the reflection schema is not assumed.

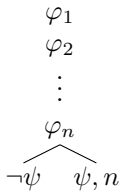
[13] See (Pawlowski and Urbaniak, 2017) for a more elaborate discussion of motivations and properties of BAT and CABAT.

3.4 BAT-trees

We will construct sound and complete tree-like proof system for BAT. The whole idea behind the proof system is to track the values of formulas which appear on branches by labeling devices called signatures. Some formulas appearing on a tree will be labeled with a letter n. Intuitively, the letter indicates that under the corresponding evaluation the formula has the value n.

We will say that a *formula φ occurs on a tree with a signature* iff it occurs on the tree in the form φ, n. Whenever we write *formula φ occurs on a tree* we mean it occurs without a signature and not as a subformula of another formula $\neg\varphi$.

Definition 36 (Root appropriate for Γ, ψ)**.** Let $\Gamma = \{\varphi_1, \varphi_2 \ldots \varphi_n\}$, $n \in \mathbb{N}$ be a set of formulas and ψ a single formula. By the *root appropriate* for Γ, ψ we will mean the following construction:

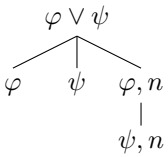


We will use syntactic rules to decompose complex formulas, extending the root appropriate for Γ, ψ:

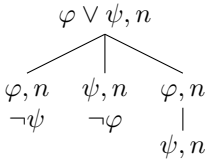
Negation 1:

Negation 2:

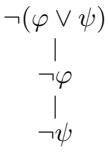
Disjunction 1:



Disjunction 2:



Disjunction 3:



Provability 1:

Provability 2:

$$\begin{array}{c} \neg \mathtt{B}\varphi \\ \hline \neg\varphi \quad \varphi, n \end{array}$$

Provability 3:

$$\begin{array}{c} \mathtt{B}\varphi, n \\ | \\ \varphi, n \end{array}$$

Since $\rightarrow, \wedge, \equiv$ are defined (in the sense of having the same matrices) as usual, we won't give their specific rules.

Definition 37 (Tree appropriate for Γ, φ)**.** By a *BAT-tree appropriate for* Γ, φ we will mean any construction that starts with the root appropriate for Γ, φ and is generated by the set of rules defined above.

Definition 38 (Full BAT-tree)**.** We say that a *BAT-tree is full* if it is not possible to apply any rule to extend the tree further.

Definition 39 (Closed branch)**.** Any path from the root is a branch of a BAT-tree. *A branch b is closed* iff for some formula φ, φ and $\neg\varphi$ occur on it, or for some formula φ it occurs on it both with and without a signature.

By the left/right root extension we will mean any path which goes down using the left/right path from the root.

Definition 40 (Closed tree)**.** A *BAT-tree is closed* iff all of its branches are closed. If at least one branch of a tree is open, the tree is open.

Note that for Γ, φ there are many different trees, depending what was the order of the rules that we applied. In this case, trees are finite, so they are order-invariant: either all of them are closed or all of them are open.

Fact 41 (Order invariance)**.** *If one tree t appropriate for Γ, φ is closed (open) so are all of them.*

Proof. Indirect. Suppose that the theorem does not hold, let t_0, t_1 be two trees appropriate for Γ, φ where t_0 is open and t_1 is closed. Let b be an open branch on t_0. This branch is constructed by a series of rules. Note, that some subset of these rules must have been applied to a certain branch b_1 on t_1. Either it was a proper subset then b is an extension of b_1 and b_1 is closed, so we have a contradiction, or both branches are generated by the same set of rules. If this is the case then again, since b_1 is closed, b so must be. □

Definition 42 (BAT consequence relation)**.** $\Gamma \vdash_B \varphi$ iff a full BAT-tree appropriate for Γ and φ is closed. By $\Gamma \nvdash_B \varphi$ we mean that BAT-tree appropriate for Γ and φ is open.

Now, we will prove that the above proof system is sound and complete with respect to BAT. We will start with some definitions and notational conventions.

Definition 43 (Faithfulness)**.** We say that an *evaluation e is faithful to a branch b* iff for all formulas φ occurring on the branch, if φ occurs without a signature then $e(\varphi) = 1$ and if φ occurs with a signature $e(\varphi) = n$.

Suppose that we have a branch b of some BAT-tree and we apply some rule to b. If the rule generates one extension with a formula φ or φ, n, we will abbreviate it as b^c, φ or b^c, φ, n. If the rule generates two or three extensions we will use b^l (or b^r) to refer to the left extension (or to the right one). In case where we have three extensions we will use b^l, b^c, b^r.

Lemma 2. Let e be a BAT-evaluation and b a branch in a BAT-tree. If e is faithful to b, then for any rule that can be applied to b, there is an extension b' of b such that e is faithful to b'.

Proof. The proof is by cases. Suppose that e is a faithful evaluation to a branch b up to the point where a formula φ occurs.

It is sufficient to check that after the application of each rule to b, the rule generates at least one extension of b which preserves faithfulness.

We will start with rules for negation:

Negation 1,2:
If $\varphi = \neg\neg\psi$, we have just one extension: b^c, ψ. The assumption implies that $e(\neg\neg\psi) = 1$. By the matrix of negation, $e(\psi) = 1$ as desired.

If $\varphi = \neg\psi, n$ we have just one extension: b^c, ψ, n. By the assumption $e(\neg\psi) = n$ and by the matrix for negation, $e(\psi) = n$.

Now consider the clauses for disjunction.

Disjunction 1:
Let $\varphi = \psi \vee \chi$. If an evaluation e is faithful to b, then either $e(\psi) = e(\chi) = n$, or $e(\varphi) = 1$, or $e(\chi) = 1$. If the evaluation e fulfills the first condition then the evaluation is faithful to b^r, if the second one is faithful to b^l, and in the third case to b^c.

Disjunction 2:
Suppose $e(\psi \vee \chi) = n$ and e is faithful to b. Then, either $e(\psi) = e(\chi) = n$, or $e(\psi) = n$ and $e(\chi) = 0$, or $e(\psi) = 0$ and $e(\chi) = n$. In the first case e is faithful to b^r, in the second e is faithful to b^l, and in the third case to b^c.

Disjunction 3:
Suppose that $e(\neg(\psi \vee \chi)) = 1$ and e is faithful to b. Then, by the matrix for negation and disjunction, $e(\psi) = 0$ and $e(\chi) = 0$ which implies that $e(\neg\psi) = 1 = e(\neg\chi) = 1$, which shows that e is faithful to b^c.

The last set of rules is the set of rules for provability.

Provability 1:
Let $\varphi = \mathtt{B}\psi$. By the assumption we only have one extension of b with ψ. By the assumption and the matrix for $\mathtt{B}$ we know that $e(\psi) = 1$ as desired.

Provability 2:
Let $\varphi = \neg\mathtt{B}\psi$. By the assumption $e(\neg\mathtt{B}\psi) = 1$. Thus, by the matrix of negation and $\mathtt{B}$, $e(\psi) = 0$ or $e(\psi) = 1$. In the first case b^l is faithful to e, and in the second case b^r preserves faithfulness.

Provability 3:
Let $\varphi = \mathtt{B}\psi, n$. By the assumption we only have one extension of b with ψ, n. By the assumption and the matrix for $\mathtt{B}$ we know that $e(\psi) = n$ as desired. □

Theorem 44 (Soundness)**.** *If* $\Gamma \vdash_B \varphi$ *then* $\Gamma \Vdash \varphi$.

Proof. Assume for contradiction that $\Gamma \vdash_B \varphi$ and $\Gamma \not\vdash \varphi$.

Consider a full tree appropriate for Γ and φ. Since $\Gamma \not\vdash \varphi$, there is an evaluation e such that $e(\psi) = 1$ for all $\psi \in \Gamma$ and $e(\varphi) \neq 1$. Suppose $\varphi \in \Gamma$. Then we have a contradiction since $e(\varphi) = 1$ and $e(\varphi) \neq 1$. Let $\varphi \notin \Gamma$. Since $e(\varphi) = 0$ or $e(\varphi) = n$, e is faithful to either the left or the right extension of the root. By induction applications of Lemma 2, there is a branch b extending the root such that e is faithful to it. Since $\Gamma \vdash_B \varphi$, we know that the branch b is closed. Thus, for some formula ψ, either ψ and $\neg\psi$ occur on it or for some formula ψ, ψ occurs with and without a signature. In the first case $e(\psi) = 1 = e(\neg\psi)$ and in the second case $e(\psi) = n$ which leads to inconsistency. □

Now we will proceed to the strong completeness of the above proof system (for finite sets[14] of formulas).

Definition 45 (Evaluation induced by b)**.** Let b be an open branch. We will say that an evaluation e is induced by b iff

- For all propositional variables p occurring without signature $e(p) = 1$,
- if $\neg p$ occurs on b then $e(p) = 0$,
- if p or $\neg p$ occurs on the branch b with signature, $e(\neg p) = e(p) = n$.

Now we will prove the completeness theorem. We will start with the following lemma:

Lemma 3. Let a branch b be open and complete. Let E be the set of evaluations induced by b. Then there is a BAT-evaluation in $e \in E$ such that:

- if φ occurs on b without a signature then $e(\varphi) = 1$,
- if $\neg\varphi$ occurs without a signature then $e(\varphi) = 0$,
- if φ or $\neg\varphi$ occurs with a signature then $e(\varphi) = n$.

Proof. Take $e \in E$ and proceed by induction on the complexity of φ. If φ is a propositional variable then we are done by definition of e being induced by b. If φ is a complex formula it has one of the forms: $\neg\psi, \psi \vee \chi, \mathtt{B}\psi$, with or without signature for some formulas ψ, χ for which the lemma already holds. We will divide this proof in two cases depending on whether φ occurs with a signature or without.

Case 1: without signature

1. Negation: suppose that $\varphi = \neg\psi$. By the induction hypothesis $e(\psi) = 0$, so by the matrix for negation $e(\neg\psi) = 1$.

   If $\varphi = \neg\neg\psi$, then the rule for double negation elimination must have been applied to obtain ψ on b. By the induction hypothesis $e(\psi) = 1$, hence $e(\neg\psi) = 0$.

[14] Since BAT and CABAT are compact, we will be interested only in finite sets of premises.

2. Disjunction: $\varphi = \psi \vee \chi$, by the completeness of the branch, one amongst ψ, χ or both ψ, n and χ, n is on the branch. In the first two cases, by the induction hypothesis $e(\psi) = 1$ or $e(\chi) = 1$, both implying that $e(\psi \vee \chi) = 1$. In the third case, by the matrix for disjunction there is an evaluation $e \in E$ such that $e(\psi \vee \chi) = 1$.

   If $\varphi = \neg(\psi \vee \chi)$, by the induction hypothesis and completeness of b, $e(\psi) = e_v(\chi) = 0$. Thus, $e(\varphi) = 0$.

3. Provability: If $\varphi = \mathtt{B}\psi$, then by the completeness of the branch we know that ψ is on the branch. By the induction hypothesis $e(\psi) = 1$ so by the matrix for $\mathtt{B}$, we have $e(\mathtt{B}\psi) = 1$.

   If $\varphi = \neg\mathtt{B}\psi$, then by the completeness of the branch we know that either $\neg\psi$ or ψ, n is on the branch. By the induction hypothesis either $e(\neg\psi)=1$ thus $e(\psi) = 0$, so $e(\neg\mathtt{B}\psi) = 1$ or $e(\varphi) = n$ and by the matrix for $\mathtt{B}$, we know that $e(\neg\mathtt{B}\varphi) = 1$.

Case 2: with signature

1. Negation: $\varphi = \neg\psi, n$. By the induction hypothesis and the matrix for negation, $e(\psi) = n$, which implies that $e(\neg\psi) = n$, as required.

   If $\varphi = \neg\neg\psi, n$. Then at some point, since the tree is complete, a rule for double negation must have been applied, so we have both: ψ and $\neg\psi$ with a signature on the branch. By the induction hypothesis $e(\neg\psi) = n$ as required.

2. Disjunction: $\varphi = \psi \vee \chi, n$ and since b is complete, either ψ, n and $\neg\chi$ or χ, n and $\neg\psi$ or both ψ and χ occur on the branch with a signature. By the induction hypothesis either $e(\psi) = n$ and $e(\neg\chi) = 1$, or $e(\chi) = n$ and $e(\neg\psi) = 1$, or $e(\psi) = n = e(\chi)$. In the first two cases $e(\psi \vee \chi) = n$ as required. In the third, by the matrix for disjunction there is at least one evaluation in E such that $e(\psi \vee \chi) = n$.

   If $\varphi = \neg(\psi \vee \chi), n$, by the completeness of the branch, $\psi \vee \chi, n$ occurs on the branch. By the induction hypothesis and the matrix for disjunction, $e(\psi \vee \chi) = n$, which implies that $e(\neg(\psi \vee \chi)) = n$.

3. Provability: If $\varphi = \mathtt{B}\psi, n$, then by the completeness of the branch ψ, n is on the branch. By the induction hypothesis $e(\psi) = n$, so we can find an evaluation for which $e(\mathtt{B}\psi) = n$.

□

Theorem 46 (Completeness)**.** *Let Γ be a set of propositional formulas, and ψ a formula. If $\Gamma \models \psi$ then $\Gamma \vdash_B \psi$.*

Proof. By contraposition. Suppose $\Gamma \nvdash_B \psi$. By definition, there is a complete open tree appropriate for Γ and ψ. Let b be the open branch in the tree. By Lemma 3 there is an evaluation e induced by b such that $e(\varphi) = 1$ for all $\varphi \in \Gamma$ and either $e(\psi) = n$, or $e(\neg\psi) = 1$ and hence $e(\psi) = 0$, depending on whether b starts with the right or the left root. In both cases we have a partial evaluation which shows that $\Gamma \not\models \psi$. □

3.5 Filtered trees

Suppose that we have a complete BAT-tree appropriate for some Γ, φ. We will devise a procedure for eliminating some of the branches in the tree in order to construct a proof system for CABAT.

Definition 47 (Filtered branch)**.** Let b be a complete open branch in a BAT-tree. We will say that b is filtered iff for all formulas φ, ψ on b the following hold:

1. If φ is a classical tautology, then it doesn't appear with a signature or in a negated form on b.
2. If φ is a classical countertautology, it doesn't appear with a signature or in an unnegated form on b
3. If φ, ψ are classically equivalent then they appear in the same form: either both with signatures, or both negated, or both in the standard form: φ, ψ.

Definition 48 (CABAT-tree)**.** By a CABAT-tree we mean any BAT-tree whose open not filtered branches are deleted.

By definition, any closed branch in a BAT-tree is not filtered.

Definition 49 (Open CABAT-tree)**.** We say that a CABAT-tree is open iff it contains an open branch b. Otherwise the tree is closed.

We use the symbol $\Gamma \vdash_c \varphi$ to denote the fact that any full CABAT-tree appropriate for Γ, φ is closed.

In order to prove completeness and soundness of the CABAT-consequence relation with respect to CABAT-trees we will use an alternative, equivalent formulation of CABAT by means of filtration of a set of evaluations:[15]

Definition 50 (CL-filtered evaluations)**.** Let CL be classical propositional logic. We say that a BAT-evaluation e is CL-filtered just in case the following conditions hold:

1. For any two formulas φ, ψ, if $\models \varphi \equiv \psi$ then $e(\varphi) = e(\psi)$,
2. For any CL-tautology φ, $e(\varphi) = 1$,
3. For any CL-countertautology φ, $e(\varphi) = 0$.

We use the symbol $\models_{fcl}$ to denote a consequence relation defined by preservation of 1 in the CL-filtered set of BAT-evaluations.

Fact 51. $\Gamma \nvdash_C \varphi$ *iff* $\Gamma \models_{fcl} \varphi$ *and any CABAT-evaluation is a CL-filtered evaluation.*

Proof. The proof can be found in (Pawlowski and Urbaniak, 2017). □

[15] See: (Pawlowski and Urbaniak, 2017).

Theorem 52. *For any finite set of formulas* Γ *and a formula* φ, $\Gamma \vdash_c \varphi$ *iff* $\Gamma \vdash\!\!\!\!\!\!\sim_C \varphi$.

Proof. $\Rightarrow$: We will argue by contraposition. Suppose that $\Gamma \not\vdash\!\!\!\!\!\!\sim_C \varphi$. We have to show that $\Gamma \nvdash_c \varphi$. By the definition of $\not\vdash\!\!\!\!\!\!\sim_C$ we know that there is a CABAT-evaluation e which assign 1 to all formulas in Γ and either 0 or n to φ. Note that any CABAT-evaluation is also a BAT-evaluation. By the completeness of the proof system $\Gamma \nvdash_B \varphi$. So there is an open branch in a BAT-tree that corresponds to e. We will argue that e is filtered, at the same time showing that it is not the case that $\Gamma \nvdash_c \varphi$.

We know that e is a CABAT-evaluation, so by fact 51 it is a filtered BAT-evaluation. It is easy to see that the conditions of a filtered branch correspond to conditions of filtered evaluation and since the branch is generated by filtered evaluation it has to be filtered as well. In other words there is at least one open filtered branch, thus $\Gamma \nvdash_c \varphi$.

$\Leftarrow$: We will argue again by contraposition. Suppose that $\Gamma \nvdash_c \varphi$, we will show that $\Gamma \not\vdash\!\!\!\!\!\!\sim_C \varphi$. By the assumption we know that there is an open filtered branch b on a tree. Take an evaluation e induced by an open filtered branch b. We will argue that this BAT-evaluation is also a CABAT-evaluation. Suppose that e is not in the set of filtered BAT-evaluation. Then it has to invalidate one of the filtration conditions. We can assume that e is a partial evaluation restricted only to the formulas on the tree, since it is a trivial matter to extend from there the partial evaluation into a full one.

Suppose that for some classical tautology φ, $e(\varphi) \neq 1$. Then either φ is on the branch with a signature or in a negated form. In both cases it is impossible, since b is filtered.

Suppose that for some classical countertautology φ, $e(\varphi) = 1$. Then φ appears on the branch in unnegated form, but this is impossible since, b is filtered.

Suppose that for some two classically equivalent formulas φ, ψ, $e(\varphi) \neq e(\psi)$. Since b is filtered and e is induced by b, we know that the form φ, ψ appear on the branch is the same: either both are negated, both are without signature or both are with signature. In all the cases evaluation e assigns to these formulas the same values. □

Chapter 4

Tree-like proof systems for finitely-many valued deterministic and non-deterministic consequence relations[1]

Abstract

The main goal of this paper is to provide an abstract framework for constructing proof systems for various many-valued logics. Using the framework it is possible to generate strongly complete proof systems with respect to any finitely valued deterministic and non-deterministic logic. I provide a couple of examples of proof systems for well-known many-valued logics and prove the completeness of proof systems generated by the framework.

4.1 Motivations

In this paper I will present an abstract framework for constructing tree-like (tableaux) proof systems for finitely-many valued deterministic and non-deterministic consequence relations.[2] Non-deterministic logics were constructed[3] by Avron and Lev (2005) as a generalization of many-valued deterministic logics. Non-deterministic logics are characterized by means of non-deterministic matrices. Roughly speaking, in such matrices an interpretation may ascribe sets of values instead of a single value. This means that these logics are not truth-functional. In their paper, Avron and Lev constructed an abstract framework for sequent-based proof systems.

I will provide a recipe for constructing tree-like (or tableaux) proof systems for non-deterministic semantics.[4] One advantage of this systems over sequent-based

[1] I am the single author of the paper and the paper is submitted.

[2] These proof systems are in the spirit of (Carnielli, 1987). For a very nice introduction to this method see (Priest, 2001).

[3] To be fair, Quine (1974) suggested something similar to a non-deterministic matrix.

[4] Roughly speaking, by non-deterministic semantics I mean a semantics in which the value of a complex formula is not uniquely determined by the values of its subformulas.

systems is that there is a nice mechanical procedure for finding a counter-model for any invalid inference. In this setting it is possible to keep track of values of formulas that appear on a tree. This allows to find valuations under which the premises have the designated value and the conclusion doesn't. Tree-like proof systems are also easy to handle for both humans and computers and they have a nice visual representation.

Moreover, it seems that some logics that use intensional operators such as modal logics or some paraconsistent logics despite not having finitely-many valued deterministic semantics, may have finite non-deterministic semantics.[5]

There is a very nice and general algebraic proof framework for non-deterministic consequence relations called *The method of Polynomial Ring Calculus*.[6] The method is mainly used in automated theorem proving. Since in that context the user-friendliness of the method is not important, unsurprisingly, the proof-system is not easy to work with. On the other hand, the tableaux method is known to be very user-friendly and pedagogically interesting, and as such deserves attention.

In the first section, I will set the notation and some necessary definitions. The second section is a presentation of the framework. In the third section I will show a couple of examples of proof systems generated using the framework for various well-known logics. The last section is a sketch of a strong completeness proof.

4.2 Technical preliminaries

Let $\mathcal{L}$ be a propositional language understood as a set of propositional variables $W = \{p, q, \ldots\}$ closed under functors from the set

$$F = \{\circ_1^1, \circ_2^1, \ldots, \circ_1^2, \circ_2^2, \ldots, \circ_m^n\}$$

where the upper script is the arity of a given functor. I will use Greek letters $\varphi, \psi, ...$ as meta-variables for formulas. I will use capital Greek letters to denote sets of formulas.

Definition 53 (n-matrix)**.** An *n-matrix* for a propositional language $\mathcal{L}$ is a triple $\mathtt{M} = \langle T, D, O\rangle$, where:

- T is a non-empty set of truth values.
- $\emptyset \neq D \subseteq T$ is a set of designated values.
- O is a set of functions, which for any n-ary functor $\circ^n$ of the language contains a corresponding n-ary function $\overline{\circ^n}$, $\overline{\circ^n} : T^n \mapsto 2^T \setminus \{\emptyset\}$.

It is quite easy to see that the above notion is a generalization of deterministic matrices. Set O consists of functions which assign sets of possible values to complex formulas, given assignments of values to their components. In the case of deterministic matrices, these sets are singletons. In a proper n-matrix

[5] Some logicians are interested in finding non-deterministic semantics for such logics (Coniglio and Peron, 2014; Coniglio et al., 2015; Omori and Skurt, 2016).

[6] See: (Carnielli and Matulovic, 2015).

functions from O can assign non-empty sets of values, including those which are not singletons. In such a case every valuation picks exactly one value of a formula from the sets of possible values ascribed to it by an appropriate function from O.

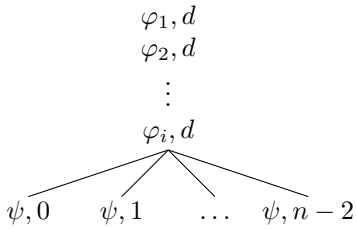
Definition 54 (Valuation)**.** A *valuation* $v : \mathcal{L} \mapsto T$ in an n-matrix $\mathtt{M}$ is a function such that for any functor $\circ^n$ and any sequence of formulas $\varphi_1, \varphi_2 \ldots, \varphi_n$, $v(\circ^n(\varphi_1, \varphi_2, \ldots, \varphi_n)) \in \overline{\circ^n}(v(\varphi_1), v(\varphi_2), \ldots, v(\varphi_n))$. A *valuation* v *satisfies* φ in $\mathtt{M}$ ($v \models_{\mathtt{M}} \varphi$) iff $v(\varphi) \in D$. We say that φ *follows from* Γ [$\Gamma \models_{\mathtt{M}} \varphi$] iff every valuation that satisfies all elements of Γ satisfies φ. An n-matrix $\mathtt{M}$ is k-valued if $|T| = k$.

Definition 55 (Consequence relation)**.** A *consequence relation* $\models$ is characterized by an n-matrix $\mathtt{M}$, iff for any Γ, φ it is the case that $\Gamma \models \varphi$ iff $\Gamma \models_{\mathtt{M}} \varphi$. We will say that a consequence relation is n-valued iff it is characterized by an n-valued n-matrix.

4.3 Trees

Suppose we have an n-valued non-deterministic logic $\models_n$ whose semantics is given by an n-matrix $\mathtt{M}$, where $T = \{0, 1, \ldots, n-2, d\}$ and d is its designated value.[7]

Definition 56. (Root appropriate Γ, ψ) Let $\Gamma = \{\varphi_1, \varphi_2, \ldots, \varphi_i\}$, be a set of formulas and ψ a single formula. By the *root appropriate* for Γ, ψ we mean the following construction:



The idea here is to list all the premises and ascribe to them index d. Roughly speaking, it means that under all the valuations corresponding to this root all premises have the designated value. Next, we split the construction into $n-2$ branches. On each branch we list the conclusion with different indices. These indices represent all non-designated values. Intuitively, each branch corresponds to any valuation under which all the premises and the conclusion values are its own indices determined by the branch.

The root is further extended into a tree by means of syntactical rules appropriate for a given logic. Before defining what a syntactical rule is we need to introduce some additional notions.

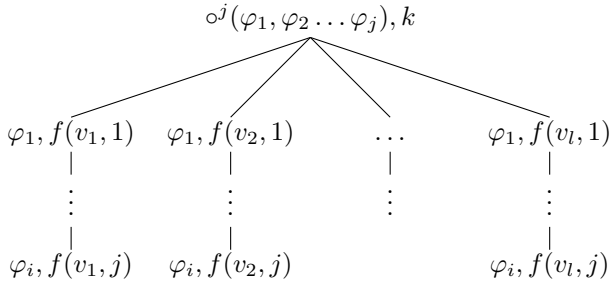
[7] We stick to the case where there is only one designated value. In order to deal with a logic whose matrix has more designated values, the validity of an inferences is checked by constructing a tree for each combination of values under which all the premises have labels corresponding to different designated values and the conclusion's label corresponds to different non-designated values. So for a reasoning with m premises in a logic with d designated values one has to construct m^d different trees. If all of them are closed, then the inference is valid. If at least one is open, it is not.

Definition 57 (i-th combination)**.** Let φ be a formula. The set of all possible valuations under which φ has value $i \in T$ is called *the i-th combination* of φ and is abbreviated as $CO_i(\varphi)$. Let $\circ^j(\varphi_1, \ldots, \varphi_j)$ be a formula built from $\circ^j$ and $\varphi_1 \ldots, \varphi_j$, and suppose that $v \in CO_i(\circ^j(\varphi_1, \ldots, \varphi_j))$. By $f(v, m)$, where $0 < m < j + 1$, we mean the value ascribed to φ_m by the valuation[8] v.

Definition 58 (Syntactical rule for $\circ^j, k$)**.** Let $\mathtt{M}$ be an n-matrix and

$$CO_k(\circ^j(\varphi_1, \varphi_2, \ldots, \varphi_j)) = \{v_1, v_2, \ldots, v_l\}.$$

By a *syntactical rule* for $\circ^j(\varphi_1, \varphi_2, \ldots, \varphi_j), k$ appropriate for an n-matrix $\mathtt{M}$ we mean the following construction:



Definition 59 (Functor described by a set)**.** Let $\mathtt{M} = \langle T, D, O \rangle$ be an n-matrix and $\circ$ be a functor from the language of the matrix. We say that a set $A_\circ$ of syntactical rules *describes* $\circ$ iff for any $t \in T$, $A_\circ$ contains a syntactical rule $\circ, t$ appropriate for $\mathtt{M}$ and nothing else is in $A_\circ$.

Definition 60 (Matrix described by a set)**.** Let an n-matrix $\mathtt{M}$ be given in the language $\mathcal{L}$, where $\circ_1, \ldots, \circ_j$ are all functors available in the language. We say that such a matrix is *described by a set of syntactical rules* A iff $A = \bigcup_{(0<i<j+1)} A_i$ where A_i describes the i-th functor in the language of n-matrix $\mathtt{M}$.

Suppose that a consequence relation is given by an n-matrix $\mathtt{M}$. In order to check any given inference in our framework we start with a root appropriate for it and then extend it into a tree by a set of syntactical rules describing $\mathtt{M}$. Next, in order to see if an inference is indeed valid in a consequence relation described by $\mathtt{M}$, we need to introduce some additional definitions.

Definition 61 (Tree appropriate for Γ, ψ)**.** A *tree is appropriate* for Γ, ψ if it starts with a root appropriate for Γ, ψ.

Definition 62 (Complete tree)**.** We say that a *tree is complete* iff it is impossible to extend the tree further using syntactical rules appropriate for a given logic.

Definition 63 (Closed branch)**.** We say that a *branch on a tree is closed* iff for some formula φ occurring on the branch it occurs with two different numbers (values).

Definition 64 (Closed tree)**.** A *tree is closed* iff all of its branches are closed.

[8]Note that we restrict our attention to partial valuations. We are only interested in the values of all subformulas. That is why, i-th combination is finite.

Definition 65 (Valid inference)**.** We say that an inference from Γ to φ is *valid on a tree* iff the tree starts with a root appropriate for Γ, φ and is complete and all of its branches are closed. We will denote this by $\Gamma \vdash \varphi$. Moreover if a tree was constructed only by using rules from a set A, we will denote this by $\vdash_A$.

4.4 A handful of examples

In this section we will show how the framework works.[9] We will give four examples of proof systems generated by our framework. Let's start with Weak Kleene Logic.[10]

4.4.1 Deterministic examples

In our case we will be interested in a propositional language with three functors $\neg, \vee, \wedge$. In Weak Kleene Logic[11] these are characterized by the following truth-tables:

$\neg$	φ
0	1
n	n
1	0

$\wedge$	0	n	1
0	0	n	0
n	n	n	n
1	0	n	1

$\vee$	0	n	1
0	0	n	1
n	n	n	n
1	1	n	1

By $\models_{Kl} \varphi$ we mean that any valuation respecting the above truth-tables ascribes value 1 to φ. By $\Gamma \models_{Kl} \varphi$ we mean that any valuation v such that for all $\psi \in \Gamma$ $v(\psi) = 1$ ascribes value 1 to φ.

The above characterization of Weak Kleene Logic can be easily rephrased in terms of an n-matrix.[12] Let A_{Kl} be a set of rules describing Kleene logics. It means that A_{Kl} consists of:

Negation 1:
$$\begin{array}{c} \neg\varphi, 0 \\ | \\ \varphi, 1 \end{array}$$

Negation 2:
$$\begin{array}{c} \neg\varphi, 1 \\ | \\ \varphi, 0 \end{array}$$

Negation 3:
$$\begin{array}{c} \neg\varphi, n \\ | \\ \varphi, n \end{array}$$

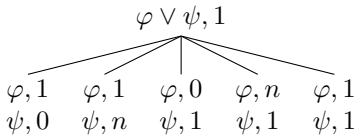
[9] Note that the results presented for deterministic consequence relations are not original. They are here for pedagogical reasons. For an extensive study of tableaux methods see (Baaz et al., 1996; Fitting, 2003).

[10] This logic sometimes goes by the name of Bochvar Logic (Bochvar, 1939). One of the main application of this logic is in truth theories (Horsten, 2011; Halbach, 2011).

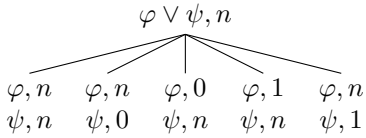
[11] Known also by the name Bochvar Internal Logic. See (Bergmann, 2008) for an extensive overview of many-valued logics.

[12] Observe that truth-tables define interpretations of functors, the designated value is 1 and $T = \{0, n, 1\}$.

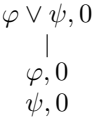
Disjunction 1:



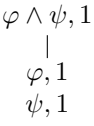
Disjunction 2:



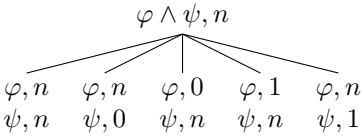
Disjunction 3:



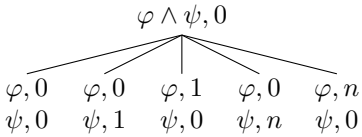
Conjunction 1:



Conjunction 2:



Conjunction 3:



It is rather easy to see that the following system is sound (since all the abstract rules are based on truth-tables). A completeness proof is in the last section.

We will proceed to another example Łukasiewicz logic (**Ł3**) (Łukasiewicz, 1970, 1988). This case is quite similar to the previous one with one difference,[13] some functors have different truth-tables:

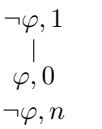
¬	φ
0	1
n	n
1	0

∧	0	n	1
0	0	0	0
n	0	n	n
1	0	n	1

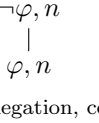
∨	0	n	1
0	0	n	1
n	n	n	1
1	1	1	1

Similarly we will use $\models_{Lk} \varphi$ to state that φ is Łukasiewicz-tautology and $\Gamma \models_{Lk} \varphi$ to denote its consequence relation (defined as a preservation of value 1). Thus, A_{Lk} contains the following rules:

Negation 1:



Negation 2:



[13] This logic has exactly the same truth-values for negation, conjunction and disjunction as strong Kleene logic. The only difference between these two lies in matrix for implication.

Negation 3:

$$\begin{array}{c} \neg\varphi, 0 \\ | \\ \varphi, 1 \end{array}$$

Disjunction 1:

$$\begin{array}{ccccc} & & \varphi \vee \psi, 1 & & \\ \swarrow & \swarrow & \downarrow & \searrow & \searrow \\ \varphi, 1 & \varphi, 1 & \varphi, 1 & \varphi, n & \varphi, 0 \\ \psi, 1 & \psi, n & \psi, 0 & \psi, 1 & \psi, 1 \end{array}$$

Disjunction 2:

$$\begin{array}{ccc} & \varphi \vee \psi, n & \\ \swarrow & \downarrow & \searrow \\ \varphi, n & \varphi, n & \varphi, 0 \\ \psi, n & \psi, 0 & \psi, n \end{array}$$

Disjunction 3:

$$\begin{array}{c} \varphi \vee \psi, 0 \\ | \\ \varphi, 0 \\ \psi, 0 \end{array}$$

Conjunction 1:

$$\begin{array}{c} \varphi \wedge \psi, 1 \\ | \\ \varphi, 1 \\ \psi, 1 \end{array}$$

Conjunction 2:

$$\begin{array}{ccc} & \varphi \wedge \psi, n & \\ \swarrow & \downarrow & \searrow \\ \varphi, n & \varphi, 1 & \varphi, n \\ \psi, n & \psi, n & \psi, 1 \end{array}$$

Conjunction 3:

$$\begin{array}{ccccc} & & \varphi \wedge \psi, 0 & & \\ \swarrow & \swarrow & \downarrow & \searrow & \searrow \\ \varphi, 0 & \varphi, 0 & \varphi, n & \varphi, 1 & \varphi, 0 \\ \psi, 0 & \psi, n & \psi, 0 & \psi, 0 & \psi, 1 \end{array}$$

4.4.2 Non-deterministic examples

The first candidate is the paraconsistent[14] logic CLuN. This logic has both a non-deterministic and a deterministic characterization (see: Batens (1999, 2000)). We will focus on the non-deterministic one. CLuN is a two-valued logic characterized by the truth-tables of classical logic for all the functors except negation. The negation is characterized by:

$\neg$	φ
{0,1}	1
1	0

where $0, 1$ means that the negated formula can have either of the two values, so negation is strictly non-deterministic. In other words, there are valuations under which both a formula and its negation can be true. Again, it is rather straightforward what the n-matrix for CLuN looks like. The set of rules appropriate for CluN is defined as:

[14] See: (Batens, 1998, 1999; Batens and De Clercq, 2004).

Negation 1:

$$\begin{array}{c} \neg\varphi, 1 \\ \hline \varphi, 0 \quad \varphi, 1 \end{array}$$

Negation 2:

$$\begin{array}{c} \neg\varphi, 0 \\ | \\ \varphi, 1 \end{array}$$

Disjunction 1:

$$\begin{array}{ccc} \varphi \vee \psi, 1 & & \\ \hline \varphi, 1 & \varphi, 1 & \varphi, 0 \\ \psi, 1 & \psi, 0 & \psi, 1 \end{array}$$

Disjunction 2:

$$\begin{array}{c} \varphi \vee \psi, 0 \\ | \\ \varphi, 0 \\ \psi, 0 \end{array}$$

Conjunction 1:

$$\begin{array}{c} \varphi \wedge \psi, 1 \\ | \\ \varphi, 1 \\ \psi, 1 \end{array}$$

Conjunction 2:

$$\begin{array}{ccc} \varphi \wedge \psi, 0 & & \\ \hline \varphi, 0 & \varphi, 1 & \varphi, 0 \\ \psi, 0 & \psi, 0 & \psi, 1 \end{array}$$

The last example of a proof system that we will formulate is more interesting, since it contains a modal operator B. The logic BAT was presented in (Pawlowski and Urbaniak, 2017) as a logic of *informal provability*. In this logic, informal provability is treated similarly to the notion of truth in (Kripke, 1975). It is partial in the sense that some sentences are informally provable, other are informally refutable but this division is not exhaustive. There are sentences which are neither informally provable nor informally refutable. To indicate this, BAT has three values: $0, n, 1$. The intended interpretation of 0 is *informally refutable*, 1 stands for *informally provable* and n stands for *neither*. BAT functors are characterized by the following truth-tables:

$\neg$	φ
0	1
n	n
1	0

$\vee$	0	n	1
0	0	n	1
n	n	{n,1}	1
1	1	1	1

$\wedge$	0	n	1
0	0	0	0
n	0	{0, n}	n
1	0	n	1

the provability operator is characterized by the following table:

B	φ
1	1
{0,n}	n
0	0

Since in this case we have multiple non-deterministic functors, we will split the presentation of syntactical rules for BAT into subsets. The negation in BAT is described by the following:

Negation 1:

$$\begin{array}{c}\neg\varphi, 1\\ |\\ \varphi, 0\end{array}$$

Negation 2:

$$\begin{array}{c}\neg\varphi, n\\ |\\ \varphi, n\end{array}$$

Negation 3:

$$\begin{array}{c}\neg\varphi, 0\\ |\\ \varphi, 1\end{array}$$

As one can imagine, rules for disjunction and conjuction are somehow complicated (both functors are non-deterministic[15]). For readability we will provide a minor simplification. Suppose that we have a formula $\varphi_1 \circ \varphi_2$. If it is clear from the truth-table that for $v(\varphi_1 \circ \varphi_2) = i$ it is sufficient that either $v(\varphi_1) = i$ or $v(\varphi_2) = i$ instead of listing all the options we will only list two options: either φ_1, i or φ_2, i. Thus, the non-deterministic functors of BAT are described by the following rules.

Disjunction 1:

$$\begin{array}{ccc}&\varphi \vee \psi, 1&\\ \swarrow & | & \searrow\\ \varphi, 1 & \psi, 1 & \varphi, n\\ & & \psi, n\end{array}$$

Disjunction 2:

$$\begin{array}{ccc}&\varphi \vee \psi, n&\\ \swarrow & | & \searrow\\ \varphi, n & \psi, n & \varphi, n\\ \psi, 0 & \varphi, 0 & \psi, n\end{array}$$

Disjunction 3:

$$\begin{array}{c}(\varphi \vee \psi), 0\\ |\\ \varphi, 0\\ |\\ \psi, 0\end{array}$$

Conjunction 1:

$$\begin{array}{c}\varphi \wedge \psi, 1\\ |\\ \varphi, 1\\ \psi, 1\end{array}$$

Conjunction 2:

$$\begin{array}{ccc}&(\varphi \wedge \psi), n&\\ \swarrow & | & \searrow\\ \varphi, n & \varphi, n & \varphi, 1\\ \psi, n & \psi, 1 & \psi, n\end{array}$$

Conjunction 3:

$$\begin{array}{ccc}&(\varphi \wedge \psi), 1&\\ \swarrow & | & \searrow\\ \varphi, 0 & \psi, 0 & \varphi, n\\ & & \psi, n\end{array}$$

Provability 1:

$$\begin{array}{c}\mathtt{B}\varphi, 1\\ |\\ \varphi, 1\end{array}$$

[15]Some disjunctions of two undecided claims can be informally provable whereas some other remains undecided. Similar considerations apply to conjuctions. For more details see (Pawlowski and Urbaniak, 2017).

Provability 2:

$$\begin{array}{c} \mathtt{B}\varphi, 0 \\ \hline \varphi, 0 \quad \varphi, n \end{array}$$

Provability 3:

$$\begin{array}{c} \mathtt{B}\varphi, n \\ | \\ \varphi, n \end{array}$$

4.5 Soundness and completeness

We only provided a general framework for constructing proof systems. Thus, the proofs of both soundness and completeness rely heavily on a particular characterization of a given consequence relation. That is why we will only sketch a general method of proving these results.

Suppose that we have a consequence relation L defined in terms of an n-matrix $\mathtt{M}$. Let A be a set of rules describing $\mathtt{M}$. In order to prove that $\vdash_A$ is sound and strongly complete with respect to L we need some further definitions.

Definition 66. We say that a valuation v is faithful to a branch b iff for all formulas φ occurring on the branch, if φ occurs with a label k, then $v(\varphi) = k$.

Fact 67. *Let v be an L-valuation and b a branch in a tree. If v is faithful to b, then for any rule of A that can be applied to b, there is an extension b' of b such that v is faithful to b'.*

Proof. We argue by cases. Let $\varphi = \circ_i(\varphi_1, \varphi_2, \ldots, \varphi_i)$ and suppose φ, k appears on b. Let $CO_k(\varphi) = \{v_1, \ldots, v_l\}$ for some l. Thus, the application of a rule appropriate for $\circ, i$ generates l many extensions of b. By construction, each extension is correlated with a valuation from $CO_k(\varphi)$ and these valuations constitute all the possible ways of having $v(\varphi) = k$. By the definition, since v is faithful to b, $v(\varphi) = k$, so $v \in CO_k(\varphi)$ which means that v is faithful to at least one extension of b. □

Fact 68. *If $\Gamma \vdash_A \varphi$ then $\Gamma \models_L \varphi$.*

Proof. By contraposition, suppose $\Gamma \not\models_L \varphi$. Then there is an L-valuation v, such that for any $\psi \in \Gamma$, $v(\psi) \in D$ and $v(\varphi) \notin D$. This valuation is faithful to at least one extension of a root of a tree for Γ, φ. By a multiple application of Lemma 67 we can find a valuation faithful to some branch b. It means that this branch must be open, so $\Gamma \not\vdash_A \varphi$. □

Definition 69 (Evaluation induced by b)**.** Let b be an open branch. An *L-valuation v is induced* by b iff for all propositional variables p, if they occur in the form p, k, then $v(p) = k$.

We proceed with a lemma which will enable us to sketch a proof of the completeness theorem.

Fact 70. *Let a branch b be open and complete. Let E be the set of L-valuations induced by b. Then there is an L-valuation $v \in E$ such that if φ occurs on b with a label k, then $v(\varphi) = k$.*

Proof. The proof is by induction on the complexity of φ. If φ is a propositional parameter, we are done by the definition of v being induced. Suppose that the theorem works for $\varphi_1, \varphi_2, \ldots \varphi_m$. We will show that it also works for $\circ^m(\varphi_1, \varphi_2 \ldots \varphi_m)$. Assume that $\circ^m(\varphi_1, \varphi_2 \ldots \varphi_m), j$ occurs on the branch. Since, b is complete a rule for $\circ^m, j$ must have been applied to it. Let

$$o = |CO_j(\circ^m(\varphi_1, \varphi_2 \ldots \varphi_m))|.$$

We have o many extensions of b where indexes of $\varphi_1 \ldots \varphi_m$ on each extension correspond uniquely to some valuation $v \in CO_j(\circ^m(\varphi_1, \varphi_2 \ldots \varphi_m))$. By the induction hypothesis if for some formula φ_i, $i < m + 1$, it occurs with a number h, there is a valuation v_h such that $v_h(\varphi_i) = h$. It is clear that $v_h \in CO_j(\circ^m(\varphi_1, \varphi_2 \ldots \varphi_m))$, so $v_h(\varphi) = j$. □

Fact 71. *If* $\Gamma \models_L \varphi$ *then* $\Gamma \vdash_A \varphi$.

Proof. By contraposition, suppose $\Gamma \nvdash_A \varphi$. Then there is a complete open tree for Γ, φ. Let b be an open branch on that tree. Take valuation $v \in E$ which is induced by b. Under this valuation all elements of Γ have the designated value and φ does not, meaning $\Gamma \not\models_L \varphi$. □

4.6 Conclusions

We presented a simple framework for constructing proof systems for finitely-valued deterministic and non-deterministic logics. Using the framework we can construct, for any finite n-matrix, a strongly complete proof system.

Chapter 5

Non-deterministic logic of informal provability has no finite characterization[1]

Abstract

Recently, in an ongoing debate about informal provability, non-deterministic logics of informal provability BAT and CABAT were developed to model the notion. I will take a closer look at some properties of these logics, proving a couple of negative results about the existence of various semantics for their consequence relations. The key results are that these logics lack finitely many-valued deterministic characterization and that a majority of commonly used modal semantics cannot characterize CABAT.

5.1 Motivations

In common mathematical practice mathematical claims are justified or proven in an informal way. Derivations are not stated in a proper formal language but rather in a mixture of a natural language expanded with mathematical notation. The existence of an informal proof of a mathematical statement is a good reason to take the claim to be known. This kind of provability connected to mathematical practice will be called *informal provability*. On the other hand, we have the notion of formal provability, relative to a given axiomatic theory.

The whole discussion about the notion of informal provability started with Myhill (1960) and Gödel (1986). Quite recently the discussion has been revived by philosophers (Leitgeb, 2009; Antonutti Marfori, 2010). They argue that there are differences between formal and informal provability. Amongst many the main differences are: the role of axioms and definitions in proofs of these two kinds is different, there is no clear algorithm for converting informal proofs into formal ones, informal proofs are better at convincing mathematicians and the justification between steps in both kinds of proofs is different.

[1] I am the single author of the paper and the paper is submitted (revisions).

Some systems for informal provability were proposed. Shapiro (1985) treats informal provability not as a predicate but as an operator. Provability is axiomatized by a modal logic **S4**. The resulting system is called Epistemic Arithmetic [**EA**]. **EA** and similar theories were further developed by Reinhardt (1986) and Horsten (1996, 1997, 1998).

Goodman (1984) showed that the internal logic[2] of **EA** is the same as a version of intuitionistic arithmetic called Heyting Arithmetic. This implies that informal provability as described by the system behaves similarly to the intuitionistic notion of provability. Given that informal mathematics is rather uncontroversially classical, it seems that **EA** does not do justice to the notion of informal provability.

A key difference between formal and informal provability lies in the principles of inference that these concepts validate. For both informal provability and formal provability the principle called *reflection schema*, which roughly says that if a given sentence is (informally) provable then it is true, is sound. Yet, not all instances of the above principle are provable for formal provability for the full language of arithmetic. The principle is crucial for informal provability. Unfortunately, even the straightforward strategy of adding all the instances of the above schema (retaining at the same time other natural conditions on informal provability) to an arithmetical theory fails — the resulting theory is inconsistent.[3] One way to solve this problem is to resign from classical logic in the background in favor of a newly developed non-classical logic aiming at modeling the notion of informal provability.

Recently, new logics BAT and CABAT for informal provability were developed (Pawlowski and Urbaniak, 2017).[4] I will present these logics and discuss some of their properties, showing that there is no finite deterministic semantics for BAT, which consequently means that BAT is a strictly non-deterministic logic. I also prove that the set of theorems of CABAT cannot be characterized by a finite deterministic matrix. Finally, I give an argument that CABAT understood as a consequence relation cannot be captured by commonly used modal semantics.

Before we proceed to the technicalities, let's take a brief look at the features of CABAT that make it an interesting candidate for the logic of informal provability. CABAT is a propositional modal logic, where B is a modal operator. The intended interpretation of B is *informally provable*. I will use $\vdash_C$ to denote its consequence relation.

Consider popular proofs of paradoxical sounding theorems involving the notion of provability, such as Montague's theorems and Löb's theorem. A vast majority of them can be translated to propositional modal language and the key

[2]By the internal logic in this context, we mean all formulas φ such that $\mathbf{EA} \vdash \Box\varphi$, where $\Box$ stands for informal provability operator of Epistemic arithmetic. See (Shapiro, 1985; Koellner, 2016) for details of Epistemic arithmetic.

[3] For clarity, note that it's possible to add the reflection principle for a formal provability predicate of a certain formal theory **T**. What we meant here is that it is impossible to have a provability predicate that shares the standard conditions put on formal provability and the reflection schema at the same time. In the first scenario the new predicate enriched by the reflection schema is no longer a formal provability predicate of the enriched theory.

[4]I am one of the authors of this paper. The paper is accepted and forthcoming in the *Review of Symbolic logic*.

moves in their proofs can be run in a propositional modal logic. Assume that we are working in a certain axiomatic theory **T** extending Peano Arithmetic. The usual strategy for such proofs is to use the diagonal lemma to generate a sentence λ provably equivalent to $\mathtt{Pr}_{\mathbf{T}}(\ulcorner\neg\lambda\urcorner)$ (for Montague's theorem and its dual version) or $\mathtt{Pr}_{\mathbf{T}}(\ulcorner\lambda\urcorner) \rightarrow \varphi$ (where $\mathtt{Pr}_{\mathbf{T}}$ is the standard formal provability predicate of **T**) for Löb's theorem. The rest of the proof proceeds on the propositional level. In order to emulate the application of the diagonal lemma we will add the propositional counterpart of the formula generated by the diagonal lemma to the premises, and proceed propositionally from there. The following fact together with its explanation should clarify the idea.

Fact 72. *The following hold for CABAT:*

1. $\mathtt{B}(\mathtt{B}\phi \rightarrow \phi), \mathtt{B}(\lambda \equiv (\mathtt{B}\lambda \rightarrow \phi)) \not\vdash_C \mathtt{B}\phi$.
2. $\mathtt{B}(\mathtt{B}\lambda \rightarrow \lambda), \mathtt{B}(\mathtt{B}\neg\lambda \rightarrow \neg\lambda), \mathtt{B}(\mathtt{B}(\neg\lambda) \equiv \lambda) \vdash_C \lambda \wedge \neg\lambda$.
3. $\mathtt{B}(\lambda \equiv \neg\mathtt{B}\lambda), \mathtt{B}(\lambda \rightarrow \mathtt{B}\lambda), \mathtt{B}(\neg\lambda \rightarrow \mathtt{B}\neg\lambda) \vdash_C \lambda \wedge \neg\lambda$.
4. *If* $\vdash_C \varphi$ *or* $\vdash_C \neg\phi$ *then* $\vdash_C \mathtt{B}\varphi \rightarrow \varphi$,
5. $\varphi \vdash_C \mathtt{B}\varphi$ *and* $\mathtt{B}\varphi \vdash_C \varphi$

1 shows that a translation of Löb's theorem does not hold in CABAT.[5] The first premise is a translation of the antecedent of Löb's theorem ($\mathbf{T} \vdash \mathtt{Pr}_{\mathbf{T}}(\ulcorner\varphi\urcorner) \rightarrow \varphi$) and the remaining premise is a translation of the application of the diagonal lemma to a formula $\mathtt{Pr}_{\mathbf{T}}(x) \rightarrow \varphi$. 2 says that it is not possible to extend CABAT with a translation of the *reflection schema*. The first two premises are translations of the reflection schema and the last one is a translation of the application of the diagonal lemma. Similarly, the third thing on the list states that one cannot consistently add all the instances of the provabilitation schema (if something is true, it is provable). Again, the first two premises are translations of the provabilitation schema for $\lambda, \neg\lambda$ and the last one a translation of the application of the diagonal lemma. Item 4 shows that the amount of reflection by default available in CABAT is greater than for the standard provability predicate. The reflection schema is not only restricted for theorems. The last item on the list shows that CABAT consequence relation preserves provability both ways (meaning a stronger version of NEC and CONEC are valid).

In classical first order arithmetic it is not possible to have the standard provability predicate for which either all the instances of the reflection schema or all the instances of the provabilitation schema are provable. The moral from the above theorem is straightforward. In CABAT it is still not possible but the amount of reflection that is valid for CABAT is greater. Moreover if we translate paradoxical theorems by implications not by consequence relation, then all the paradoxes are no longer valid. Unfortunately, the conditional in CABAT is quite weak so this is not a fully satisfactory solution.

[5] To be fair if we add to the premise set a formula $\mathtt{B}\neg\lambda \rightarrow \neg\lambda$ then the consequence does hold. But we are not interested here in an arbitraty instance of the reflection schema but in the translation of the standard proof of Löb's theorem. Thus, in general the reflection schema is not assumed.

In the next section of this paper I will introduce some key technical notions. Then, I will proceed with a presentation of BAT and CABAT. In section 4, I will further elaborate on some properties of these logics. In section 5 and 6 I will prove the key negative results about the existence of various types of semantics sound for either BAT or CABAT.

5.2 Technical preliminaries

Let $\mathcal{L}$ be a propositional language (understood as the set of all well-formed formulas) constructed from propositional variables $Var = \{p_1, p_2, \dots\}$, and connectives. Usually, I will be interested in the language where we have the classical Boolean connectives: $\neg, \wedge, \vee, \rightarrow, \equiv$. I will use Greek letters $\phi, \psi, ...$ as meta-variables for formulas and by Wl we will mean the set of all formulas. Suppose $\mathcal{L}$ is a language, by $\mathcal{L}_{\mathtt{B}}$ I will mean an extension of this language with a unary operator $\mathtt{B}$. In our case, we use $\mathtt{B}$ rather than a more common $\Box$ since we are mostly interested in interpretations where the operator stands for informal provability.

In order to define BAT and CABAT I will need non-deterministic matrices developed in (Avron and Lev, 2001, 2005; Avron and Zamanski, 2011).

Definition 73. A matrix for a propositional language $\mathcal{L}$ is a tuple $M = \langle T, D, O\rangle$, where T is a non-empty set of truth values, $D \subseteq T$ is the set of designated values. O is the smallest set that for any n-ary connective $\circ^n$ of the language, there is a corresponding n-ary function $\overline{\circ^n} \in O$ such that $\overline{\circ^n} : T^n \mapsto T$.

Definition 74. A valuation in M is a function $v : Wl \mapsto T$ such that for any connective $\circ^n$ and any sequence of formulas $\varphi_1, \varphi_2 \dots, \varphi_n$, $v(\circ^n(\varphi_1, \varphi_2, \dots, \varphi_n)) = \overline{\circ^n}(v(\varphi_1), v(\varphi_2), \dots, v(\varphi_n))$. For any Γ, φ a valuation v:

1. Satisfies φ in M in symbols $v \models_M \varphi$ iff $v(\varphi) \in D$.
2. Is a model of Γ ($v \models_M \Gamma$) iff v satisfies every formula in Γ.

We say that φ is M-valid iff for any valuation in M, $v \models_M \varphi$. In this case we often skip the reference to the valuation. By $\Gamma \models_M \varphi$ we mean that for any valuation v, if $v \models_M \Gamma$ then $v \models_M \varphi$. In this case we say that φ follows from Γ.

To have a description of non-deterministic logics in terms of matrices we need to modify the notion of a matrix and the notion of a valuation.

Definition 75. An n-matrix for a propositional modal language $\mathcal{L}$ is a tuple $M = \langle T, D, O\rangle$, where T is a non-empty set of truth values, $D \subseteq T$ is the set of designated values. O is the smallest set such that for any n-ary connective $\circ^n$ of the language, there is a corresponding n-ary function $\overline{\circ^n}$, $\overline{\circ^n} : T^n \mapsto 2^T \setminus \{\emptyset\}$ in O.

This notion is a generalization of the notion of a matrix for a propositional logic. In the standard deterministic setting, O assigns exactly one value for any combination of connectives where in an n-matrix it assigns a non-empty subset of D.

The role of an evaluation in non-deterministic setting is to specify a single value ascribed by O. In other words the value of a complex formula doesn't depend solely on the values of its arguments.

Definition 76. A valuation in an n-matrix M is a function $v : Wl \mapsto T$ such that for any connective $\circ^n$ and any sequence of formulas $\varphi_1, \varphi_2 \ldots, \varphi_n$, $v(\circ^n(\varphi_1, \varphi_2, \ldots, \varphi_n)) \in \overline{\circ^n}(v(\varphi_1), v(\varphi_2), \ldots, v(\varphi_n))$. A valuation v:

1. Satisfies φ in M ($v \models_M \varphi$) iff $v(\varphi) \in D$.

2. Is a model of Γ [$v \models_M \Gamma$] iff v satisfies every formula in Γ.

We will say that φ follows from Γ [$\Gamma \models_M \varphi$] iff every model of Γ is a model of φ. An n-matrix M is n-valued if $|T| = n$.

Definition 77. A consequence relation $\models$ is characterized by:

1. A matrix M, iff for any Γ, φ it is the case that $\Gamma \models \varphi$ iff $\Gamma \models_M \varphi$.

2. An n-matrix M iff $\Gamma \models \varphi$ iff $\Gamma \models_M \varphi$ for any Γ.

The above definitions are stated in a general format. Mostly through this paper we will limit our interest to the classical propositional language and its extension with a provability operator.

5.3 BAT and CABAT

CABAT and BAT were introduced in (Pawlowski and Urbaniak, 2017) as *logics for informal provability*. In order to define CABAT we have to start with the definition of BAT (BAT comes from Being an Absolute Theorem).[6]

BAT is a three-valued $(0, n, 1)$ non-deterministic logic. The intended interpretation of 0 is *informally refutable*, 1 stands for *informally provable* and n for *neither*. The idea behind the three-valued semantics relies on the intuition that in mathematical practice we can distinguish three different sets of mathematical claims: those which have an informal proof, those which have an informal refutation, and undecided ones.

Prima facie, we can distinguish two different interpretations of informal provability. According to the first one, informal provability is time dependent. For instance, currently a mathematical claim may be informally undecidable, but later some mathematician can prove it. It will no longer be independent. Thus, the status of a sentence can change over time.

According to the second interpretation, informal provability is time independent: some are simply provable, some are simply refutable, and some are simply independent. The status of mathematical sentences is fixed and cannot change.

If we stick to the second interpretation then for some undecidable mathematical sentences we may be able to prove that their conjunction is informally

[6]The etymology of "CABAT" is a bit more complex. The letter C in the name stands for the closure condition and the letter A, for algebraic conditions that turned out to be redundant.

refutable.[7] For instance, consider the Continuum Hypothesis and its negation. For some other, their conjunction may remain informally undecidable. For instance, take two sentences: one expressing the consistency of ZFC and the other expressing the Continuum hypothesis. The reasoning carries over to disjunction. Some disjunctions of undecided sentence may be undecided and some others informally provable. It seems, informal provability is not truth-functional. This is the main motivation for using non-deterministic matrices.[8]

Let $M = \langle T, D, O\rangle$ be an n-matrix where $T = \{0, n, 1\}$, $D = \{1\}$ and O is the set of functions defined by the following tables:

$\neg$	ϕ
0	1
n	n
1	0

$\vee$	0	n	1
0	0	n	1
n	n	$\{n, 1\}$	1
1	1	1	1

$\wedge$	0	n	1
0	0	0	0
n	0	$\{0, n\}$	n
1	0	n	1

$\rightarrow$	0	n	1
0	1	1	1
n	n	$\{n, 1\}$	1
1	0	n	1

$\equiv$	0	n	1
0	1	n	0
n	n	$\{0, n, 1\}$	n
1	0	n	1

B	ϕ
1	1
{ 0, n}	n
0	0

By a *BAT-valuation* we mean every non-deterministic valuation which respects the above tables. The consequence relation of BAT relation is denoted by $\vdash$ and is defined in the usual manner: $\Gamma \vdash \phi$ iff for all BAT-valuations, if $v(\psi) = 1$ for all $\psi \in \Gamma$ then $v(\phi) = 1$.

We say that a BAT-valuation e respects the *closure condition* iff

> For any $\mathcal{L}_{\mathtt{B}}$-formulas $\phi_1, \phi_2, \ldots, \phi_n, \psi$ such that
>
> $$\phi_1, \phi_2, \ldots, \phi_n \models \psi,$$
>
> where $\models$ is the classical consequence relation for $\mathcal{L}_{\mathtt{B}}$, if $v(\mathtt{B}\phi_i) = 1$ for any $0 < i \leq n$, then $v(\mathtt{B}\psi) = 1$.

CABAT is the logic resulting from BAT by not considering BAT-valuations which do not respect the closure condition. The closure condition is quite important since BAT is a very weak logic, for instance disjunction in BAT is not symmetric.[9]

[7] For the first interpretation, it's impossible since the status of a mathematical claim may change. For instance, currently the Goldbach hypothesis is undecidable but it may turn out later that we will be able to prove or disprove it. Thus, a conjuction and disjuction of two independent sentences may turn out to be provable, refutable or remain undecided.

[8] To be fair, it is also a perfectly viable motivation to approach informal provability from supervaluationistic perspective. This approach and the comparison between the two approaches, however, is beyond the scope of this paper.

[9] On the other hand it is an interesting system to study since it is a starting point and system on top of which CABAT is defined.

We will use $\Gamma \vdash_C \phi$ to denote the consequence relation of CABAT. Alternatively, we can characterize CABAT by means of filtration:

Definition 78. Let $\models$ be the classical propositional consequence relation in $\mathcal{L}_{\mathtt{B}}$. We say that a BAT-valuation e belongs to the filtered set of BAT-valuations just in case the following conditions hold:

1. For any two formulas ϕ, ψ, if $\models \phi \equiv \psi$ then $v(\phi) = v(\psi)$,
2. For any tautology ϕ, $v(\phi) = 1$,
3. For any countertautology ϕ, $v(\phi) = 0$.

Let $\Gamma \models_{Lf} \varphi$ mean that for any L-filtrated BAT-valuation v such that $v(\psi) = 1$ for any $\psi \in \Gamma$ it follows that $v(\varphi) = 1$. Then the following holds:

Fact 79. *For any* Γ, φ, $\Gamma \vdash_C \varphi$ *iff* $\Gamma \models_{Lf} \varphi$

Proof. See (Pawlowski and Urbaniak, 2017). □

One may wonder why we added an additional operator B to the language. *Prima facie* it seems that this operator is defined by $\mathtt{B}\phi := (\phi \vee \neg\phi) \wedge \phi$ since both formulas have the same matrix. But this does not imply that for every valuation both formulas have the same values. The semantics is non-deterministic so there is a valuation e such that $v(\mathtt{B}\phi) = n$ and $v((\phi \vee \neg\phi) \wedge \phi) = 0$. The general phenomenon is that if two formulas have the same truth-table it does not mean that for each valuation the values of these formulas are the same. But if for any valuation the value of two formulas is the same then they have the same table.

5.4 The lack of finite deterministic characteristic

In this section we show that both BAT consequence relation[10] and the set of CABAT tautologies[11] cannot be characterized by a finite deterministic matrix. We will proceed indirectly. We assume that such a characterization exists. Next, for each natural number n we will give a general recipe, for constructing a set Γ_n and a single formula ψ_n such that any [potential] deterministic n-valued characterization of BAT validates the reasoning from Γ_n to ψ_n, but the inference is not valid in BAT.

For the latter result we employ a reasoning used by Dugundji (1940) in a modern formulation that can be found in (Coniglio and Peron, 2014).

The intuition here is that we will use the lack of symmetricity for disjunction in BAT. Let $\Gamma_n = \{p_i \vee p_i \mid 0 < i < n+2\} \cup \{p_i \vee p_j \mid 0 < i < j < n+2\}$. Formula ψ_n is defined as $\bigvee_{i<j}^{j=n+1}[(p_i \vee p_j) \wedge (p_j \vee p_i)]$.

Lemma 4. For any n, $\Gamma_n \not\vdash \psi_n$.

[10] Since BAT does not have any tautologies, we have to stick to its consequence relation.

[11] From the fact that CABAT understood as the set of tautologies cannot have a finite many-valued characterization, it follows that CABAT understood as a consequence relation cannot have deterministic n valued semantics as well.

Proof. First, since disjunction is not symmetric in BAT, for any i, j, we have $p_i \vee p_j \not\vdash p_j \vee p_i$. It is rather easy to see that for each i, j, $p_i \vee p_i, p_i \vee p_j \not\vdash p_j \vee p_i$, so $p_i \vee p_i, p_i \vee p_j \not\vdash (p_j \vee p_i) \wedge (p_i \vee p_j)$. This straightforwardly generalizes to any subset of Γ. □

Now, we will show that any (potential) n-valued deterministic logic has to validate the reasoning from Γ_n to ψ_n, so it cannot be the case that for all Γ, φ, $\Gamma \models_n \varphi$ iff $\Gamma \vdash \varphi$.

Theorem 80. *For each n, $\Gamma_n \models_n \psi_n$.*

Proof. Suppose for a contradiction that $\Gamma_n \not\models_n \psi_n$. Let e be a deterministic valuation such that for any $\varphi \in \Gamma_n$, $v(\varphi) \in D$ and $v(\psi_n) \notin D$. Formula ψ_n contains $n+1$ different propositional variables. Since the logic is n-valued, there are i, j such that $i \neq j$ and $v(p_i) = v(p_j)$.

Since this characteristic is deterministic it follows that $v(p_i \vee p_i) = v(p_i \vee p_j)$ and $v(p_j \vee p_j) = v(p_j \vee p_i)$. But $p_i \vee p_i, p_j \vee p_j \in \Gamma_n$, so $v(p_i \vee p_i), v(p_j \vee p_j) \in D$. It follows that $v(p_i \vee p_j) = v(p_j \vee p_i) \in D$. Since, $\psi_1, \psi_2 \vdash \psi_1 \wedge \psi_2$, we have $v((p_i \vee p_j) \wedge (p_j \vee p_i)) \in D$ on the assumption that the deterministic matrix under consideration captures BAT consequence. But also, for any two formulas φ_1, φ_2 we have $\varphi_1 \vdash \varphi_1 \vee \varphi_2$, and so we also have $v(\psi_n) \in D$, which contradicts the assumption. □

We shall proceed now to showing that CABAT cannot be characterized by a finite matrix. In the proof we will employ a certain version of Dugundji's formula used in (Coniglio and Peron, 2014). By $p \Rightarrow q$ we will mean the formula $\mathtt{B}(p \to p) \to \mathtt{B}(q \to p)$. For each natural number n we will construct a so-called Dugundji formula: $\mathcal{DG}_n = \bigvee_{i \neq j}(p_i \Rightarrow p_j)$, where $0 < i, j < n + 2$.

The general strategy for the proof is as follows. For a contradiction we will assume that there is an n-valued deterministic matrix characterizing CABAT $\models_n$, meaning that for any φ, $\models_n \varphi$ iff $\vdash_C \varphi$. Then we will show that any such matrix validates D_n. Then we will argue that CABAT does not validate D_n for any n.

First, we will show that any n-valued characterization of CABAT validates D_n. Suppose that $\models_n$ is such that for any φ, $\models_n \varphi$ iff $\vdash_C \varphi$.

Lemma 5. For any natural number n, $\models_n \mathcal{DG}_n$.

Proof. Take any valuation v. Since the matrix is n-valued, and the formula has $n+1$ variables, by the pigeon-hole principle at least two variables have the same value. Thus, there are i, j such that $v(p_i) = v(p_j)$. Since the characterization is deterministic, it follows that $v(p_i \Rightarrow p_j) = v(p_i \Rightarrow p_i)$. Note that $p_i \Rightarrow p_i$ is a CABAT tautology. By propositional logic, for any formulas ϕ, ψ, we have $\vdash_C \phi \vee (p_i \Rightarrow p_i) \vee \psi$. From $v(p_i \Rightarrow p_j) = v(p_i \Rightarrow p_i)$ and the last observation it follows that the matrix validates $\mathcal{DG}_n$. □

Fact 81. *For any natural number n, $\not\vdash_C \mathcal{DG}_n$.*

Proof. Consider an evaluation e, under which all propositional variables used in $\mathcal{DG}_n$ have value n. It is rather easy to see, that under this evaluation each disjunct of $\mathcal{DG}_n$ of the form $\mathtt{B}(p_i \to p_i) \to \mathtt{B}(p_j \to p_i)$ can have value n, which makes the whole big disjunction n as well, showing that it is not a CABAT tautology. □

5.5 Lack of modal semantics

In this section we will elaborate on modal semantics for CABAT. In particular, we will show that neither (normal or non-normal) Kripke[12] semantics, nor neighborhood semantics can capture CABAT understood as a consequence relation.

First, observe:

Fact 82. $\not\vdash_C \mathtt{B}(\varphi \to \psi) \to (\mathtt{B}\varphi \to \mathtt{B}\psi)$.

This means that CABAT cannot have standard Kripke semantics since axiom K is valid in all possible world models.

One may think that the lack of K in CABAT is problematic. It seems that axiom K is a correct principle for informal provability. If an implication is informally provable, then if one can prove its antecedent, one has a proof of its consequent. This is not a problem in CABAT since implication in CABAT says something more than axiom K does. Observe, $\vdash_C \mathtt{B}(\varphi \to \psi) \to (\mathtt{B}\varphi \to \mathtt{B}\psi)$ would not only inform us what happens when the antecedent has value 1, but also what happens when its value is n. For instance, if $v(\mathtt{B}(\varphi \to \psi)) = n$ then the consequent $\mathtt{B}\varphi \to \mathtt{B}\psi$ couldn't have value 0. Thus, the initial intuition supporting the validity of K, "If an implication is informally provable, then if one can prove its antecedent then one has a proof of its consequent," should be translated as $\mathtt{B}(\varphi \to \psi)$, $\mathtt{B}\varphi \vdash_C (\mathtt{B}\varphi \to \mathtt{B}\psi)$, and this inference is valid in CABAT. So, the intuition behind K is indeed captured in CABAT as well.

It seems that theorem-wise the closure condition of CABAT assures us that CABAT is closed under (Nec) and (Conec) (if $\vdash \Box\varphi$ then $\vdash \varphi$). However note that CABAT is slightly stronger than a modal logic closed under (Nec) and (Conec), as the following fact indicates:

Fact 83. *Let $\models_{NC}$ be classical propositional logic in the language with a modal operator closed under (Nec) and (Conec). Then the following is the case: for all φ, if $\models_{NC} \varphi$ then $\vdash_C \varphi$ and there is a formula λ such that $\vdash_C \lambda$ but it is not the case that $\models_{NC} \lambda$.*

Proof. We will show by induction on the number of applications of (Nec) and (Conec) that for each line i in the proof $\vdash_C \varphi_i$.

Suppose $n = 0$. It means that φ_i is a tautology of classical logic. CABAT validates modus ponens and all propositional tautologies, which means φ is valid in CABAT. Assume the theorem holds for n applications of (Nec) or (Conec). We will show that the theorem holds for $n+1$. Let $\varphi_1 \dots \varphi_k$ be an NC proof of φ, so $\varphi_k = \varphi$. Let l be the line in a proof obtained by the $n+1$-th application of (Nec) or (Conec). By the induction hypothesis, any formula at a line occurring

[12] A non-normal Kripke semantics may contain non-normal worlds, resulting in the weakening of (Nec) [if $\vdash \varphi$ then $\vdash \Box\varphi$] (Chellas, 1980).

before line l is valid in CABAT. But CABAT is closed under both (Nec) and (Conec), so a formula at line l is also valid in CABAT, for it was obtained from a formula occurring before, which is valid in CABAT, by means of either (Nec) or (Conec).

For the second part of the theorem consider $\lambda = \mathtt{B}(q \wedge \neg q) \rightarrow p$. Clearly, $\not\vdash_C \lambda$ but it is easy to see that we cannot get this by a simple application of (Nec) and (Conec). □

Recall that a non-normal Kripke frame is obtained from the standard Kripke frame by distinguishing between two kinds of worlds: normal and non-normal ones. In non-normal worlds nothing is necessary and everything is possible. Using this kind of semantics we can model logics where necessitation is not valid. Yet, every non-normal Kripke frame validates axiom K, so it cannot be an adequate semantics for CABAT.

One of the most general modal frameworks is the so-called neighborhood semantics. In this semantics, instead of an accessibility relation, we have a neighborhood function. This function fixes which propositions are necessary at which world.[13]

Definition 84. Let W be a set of worlds. A function $N : W \mapsto P(P(W))$, where $P(W)$ is the power set of W, is called a neighborhood function.

Definition 85. A tuple $\langle W, N \rangle$ is called a neighborhood frame iff W is non-empty and N is a neighborhood function.

Definition 86. A tuple $\langle W, N, v \rangle$ is called a neighborhood model iff $\langle W, N \rangle$ is a neighborhood frame, and $v : At \mapsto P(W)$ is a valuation function.

Definition 87. Let $M = \langle W, N, v \rangle$ be a neighborhood model. Let $w \in W$. Satisfaction conditions for formulas are defined in the following way:

1. $M, w \models p$ if $w \in v(p)$
2. $M, w \models \neg\varphi$ if $M, w \not\models \varphi$
3. $M, w \models \varphi \wedge \psi$ if $M, w \models \varphi$ and $M, w \models \psi$
4. $M, w \models \Box\varphi$ iff $(\varphi)^M \in N(w)$

where $(\varphi)^M = \{w \mid M, w \models \varphi\}$ is the truth-set of φ. The first three conditions are straightforward. The fourth condition states that a given formula is necessary in a given world if the set of worlds in which the formula is true belongs to a family of sets ascribed by the neighborhood function to this particular world.

Let M be a neighborhood model. We say that φ is valid in a neighborhood model M [$M \models \varphi$] iff for any point $w \in W$, $M, w \models \varphi$. If φ is valid in any neighborhood model, then we will say that it is valid in a neighborhood frame. Using validity in a frame we can define the notion of a local semantical consequence.

[13] For more information on this semantics see (Chellas, 1980).

Definition 88. A formula φ is a local semantic consequence[14] of a set of formulas Γ iff for all neighborhood models M and all points $w \in W$, if $M, w \models \Gamma$ then $M, w \models \varphi$.

Note that any local semantical consequence validates the following:

Theorem 89. *Let $\models$ be a local neighborhood semantical consequence relation. Then for any two formulas φ, ψ, if $\varphi \models \psi$ and $\psi \models \varphi$ then $\models \varphi \equiv \psi$.*

Proof. Suppose that $\varphi \models \psi$ and $\psi \models \varphi$. Suppose also by a contradiction, that $\not\models \varphi \equiv \psi$. We have two cases to consider: either there is a model $M = \langle W, N, v \rangle$ such that $M \models \varphi \wedge \neg\psi$ or $M \models \neg\varphi \wedge \psi$. Since both cases are symmetric we will only show the first one. Since $M \models \varphi$, by the assumption, we have $M \models \psi$, which is a contradiction. □

But the above is not true in CABAT:

Fact 90. *From $\varphi \blacktriangleright_C \psi$ and $\psi \blacktriangleright_C \varphi$ it does not follow that $\blacktriangleright_C \varphi \equiv \psi$.*

Proof. To see that take a valuation on which both φ and ψ have value n and assume that both $\varphi \blacktriangleright_C \psi$ and $\psi \blacktriangleright_C \varphi$. It is easy to see that under this valuation the equivalence may have value n. □

This consequence shows that its impossible to characterizes CABAT by neighborhood semantics.

[14] Let F be a family of frames. We say that φ is a *global* semantic consequence of Γ ($\Gamma \models \varphi$) iff for all frames $G \in F$, if $G \models \Gamma$ then $G \models \varphi$. For more details about this distinction see Chapter 1 of (Blackburn, 2001).

Chapter 6

Paradoxes of informal provability and many-valued non-deterministic provability logic[1]

Abstract

Paradoxes of informal mathematical provability are used to argue for the inconsistency of the notion of informal provability and for dialetheism in general. We discuss these paradoxes, their formal counterparts, and their alleged role in the defense of dialetheism. We argue that the dialetheist arguments fail, but not for the reasons put forward so far in the literature. The second part of the paper is more constructive: we introduce non-classical logics of informal provability (BAT and CABAT) and approach the paradoxes wielding this weapon.

Keywords. Informal provability, many-valued logic, non-deterministic semantics, Löb's theorem, paradoxes of provability

6.1 Formal vs Informal provability

Mathematicians justify or prove their claims in an informal way. Their *informal proofs* aren't really stated in a proper formal language, but rather in a mixture of natural language expanded with mathematical notation, and aren't as rigorous as formal proofs. Sometimes, it isn't even clear what counts as an axiom and some simple facts are said to be justified merely on the basis of mathematical insight (or intuition). Yet, the existence of an informal proof of a mathematical statement is a compelling reason to take the claim to be true (or established).

[1]This is a joint paper with Rafal Urbaniak as the first author. The paper is currently submitted.

Formal proofs, on the other hand, are formulated in a fully formalized axiomatic theory, and employ very specific formal rules of proof construction.[2]

The standard view[3] on the relation between formal and informal provability is that any informal proof is at least in principle reducible to a proper proof in an appropriate axiomatic system (usually, ZFC) (Rav, 1999; Sjögren, 2010; Antonutti Marfori, 2010). On this view, informal proofs are just sloppy, incomplete versions of formal proofs. However, some philosophers (Myhill, 1960; Horsten, 2002; Leitgeb, 2009; Antonutti Marfori, 2010) argue against this view. The standard view, they insist, does not fully explain why informal proofs are quite good at convincing mathematicians, whereas formal ones are not. They also point out that the role of axioms and definitions is quite different in both kinds of proofs and that there is no clear procedure for converting an informal proof into a formal one or for associating informal proofs with their formal counterparts (Tanswell, 2015).

While we do find those aspects of the comparison of these two notions interesting, we won't get into the issues that are usually brought up in the debate. Instead, we'd like to focus on general inference principles: are there any inference principles on which formal and informal provability disagree, that is, which hold for one, but don't hold (or shouldn't hold) for the other (and what is meant by 'hold' in this context)?

One example of a principle for which a difference seems to arise is the *reflection schema*. Roughly speaking, it says that any provable sentence is also true. On one hand, reflection is not too compelling for formal provability *as such*: to think that claims provable in an axiomatic theory are true, one has to also assume that the axioms of the theory are true and that the inference rules are truth-preserving. For instance, it is not the case that for any axiomatic system built over the standard arithmetical language, whatever is provable in that system is true in the standard model. For reflection for such a system to be true, it also has to be the case that the axioms of that system are true in the standard model (and that provability in the system preserves truth). In contrast, claims proven in informal mathematics are taken to be true in virtue of having an informal proof — for this reason, the reflection schema for *informal provability* (henceforth *informal reflection schema*) is quite compelling.

Unfortunately, due to Löb's theorem, the language of any consistent recursively axiomatizable arithmetical theory `T` containing Peano Arithmetic cannot contain a formula for which the standard Hilbert-Bernays conditions on provability (to be listed soon) and reflection for the full language of `T` hold.

[2]The whole discussion on this topic was initiated by Gödel (1953).

[3]This view is usually shared by mathematicians; for instance Enderton (1977, 10-11) says:

> It is sometimes said that "mathematics can be embedded in set theory." This means that mathematical objects (such as numbers and differentiable functions) can be defined to be certain sets. And the theorems of mathematics (such as the fundamental theorem of calculus) then can be viewed as statements about sets. Furthermore, these theorems will be provable from our axioms. Hence our axioms provide a sufficient collection of assumptions for the development of the whole of mathematics — a remarkable fact. (In Chapter 5 we will consider further the procedure for embedding mathematics in set theory.)

Also, for a bit more sophisticated version of the standard view, see (Sjögren, 2010).

Things are even worse than Löb's theorem might suggest. Myhill (1960) and Montague (1963) proved that no formal theory extending Robinson arithmetic admits a provability predicate for which the reflection schema and Nec (if $\vdash \varphi$ then $\vdash P(\ulcorner\varphi\urcorner)$) hold. So, at least *prima facie*, if we wanted to axiomatize our intuitions about informal provability over arithmetic, we might have a really hard time consistently incorporating some rather intuitive principles.

Some attempts at formulating a formal theory of informal provability have been made. Shapiro (1985) constructed a theory called *Epistemic Arithmetic* (EA) where informal provability is formalized as an operator rather than as a predicate. On this approach informal provability is governed by a modal logic S4. Shapiro defined a theorem-preserving translation v from the language of arithmetic based on intuitionistic logic (Heyting arithmetic) to the language of EA. Goodman (1984) proved that the translation is faithful and his proof was further simplified by Flagg and Friedman (1986). This theory was further developed into three directions: by considering additional principles such as *Epistemic Church Thesis* (Flagg and Friedman, 1986; Halbach and Horsten, 2000), extending the language with a truth predicate (Stern, 2015; Koellner, 2016), and a deeper analysis of the informal provability operator (Horsten, 1994, 1997; Heylen, 2013). The whole framework was further studied by Rin and Walsh (2016).

The collateral damage here is a serious limitation on expressive power that results from treating provability as an operator rather than as a predicate.[4] It is no longer possible to quantify over formulas by means of coding. Moreover, the internal logic of the operator in the theory is quite weak, because of the existence of the previously mentioned translation from intuitionistic logic.

A different approach to informal provability was proposed by Horsten (1997). The idea is simple: informal provability remains a predicate, but the set of principles holding for it is weakened to a rather small set of very uncontroversial ones. The set of intuitive inferential principles is split into two, and they're added to two arithmetical theories called *the basis* and *the main theory*. The main theory is further extended with the principle saying that if something is provable in the basis, it is informally provable. The main feature of the approach is that while reflection indeed holds for the main theory, certain other principles, such as (Nec), can be applied only to the basis.

This approach seems to be more promising but it has its own problems. There seems to be no principled way of deciding which principles should be weakened and to what extent. Stern (2015) proved that many similar systems are inconsistent. So the approach doesn't seem as promising as one might initially hope.

Recently, Pawlowski and Urbaniak (2017) proposed an alternative way to build a theory of informal provability by changing the underlying logic. The authors developed non-deterministic three-valued logics which can be used for building a formal theory of informal provability. On this approach the move to the predicate level seems viable. The translations of paradoxical theorems blocking the move to the predicate level for theories having classical logic in the background do not hold. One of the aims of this paper is to investigate how

[4]This feature is essential: one cannot extrapolate to the predicative level, for at that level one could employ the Diagonal Lemma to repeat the moves present in the proof of Montague's theorem.

quasi-paradoxical theorems can be translated into this setting and to investigate how they behave in this context.

Section 6.2 discusses paradoxes related to provability. First, in section 6.2.1 we sketch informal versions of key paradoxical arguments. In section 6.2.2 discuss their formalizations in order to see what theorems result from approaching them formally. In section 6.3 we examine Löb's theorem and its alleged applications to informal mathematics. Section 6.4 is devoted to paradoxes formulated using the notion of informal provability. Notably, in section 6.4.1 we take a closer look at a *dialetheist argument* for the inconsistency of informal mathematics. We dissect its problematic premises in sections 6.4.2 and 6.4.3. Then we move on to the constructive part: the development of a non-classical logic of informal provability. We start with lying down motivations in subsection 6.5.1 and with a sketch of the general background strategy in subsection 6.5.2. The main goal of subsection 6.5.3 is to present our non-deterministic approach where we develop two logics: BAT and CABAT. These logics are discussed further in subsection 6.5.4 and section 6.6. We conclude, in section 6.7, with an application of these logics to paradoxical arguments discussed throughout the paper.

6.2 Paradoxes of provability

6.2.1 Informal paradoxes

Here we will take a closer look at some well-known principles involving provability that have been considered problematic or paradoxical. Let's begin with the *Informal Provability Gödel* sentence: a sentence which says that it is not informally provable.

(IPG) is not informally provable. (IPG)

At least *prima facie*, (IPG) gives rise to a paradox in the vein of (Beall, 1999; Priest, 2006), where the argument is put forward to the effect that informal mathematics is inconsistent. For suppose (IPG) is false. This means it is informally provable. Then, assuming informal reflection, it is true. The assumption that (IPG) is false leads to the conclusion that it is true, which itself proves that (IPG) is true.[5] But this piece of reasoning is an informal proof, and so we have just informally proven (IPG), which means that it is informally provable. This, however, means that it is false after all. Contradiction.

Let's call sentences whose negations are informally provable *informally refutable*.[6] A close kin of (IPG) is what we'll call the *Informal Provability Liar* — a sentence that says of its own negation that it is informally provable.

(IPL) is informally refutable. (IPL)

Again, it seems that we can use (IPL) to reason to a contradiction. For suppose (IPL) is true. If that's the case, things are as it says, and so (IPL) is informally

[5] That is, we use $(\neg p \to p) \to p$. The argument can be easily run as a *reductio*.

[6] Mind your head: proving informal refutability *is not* the same as proving underivability. An informally refutable claim is *not* independent: it's rejected.

refutable, and the negation of (IPL) is informally provable. By informal reflection, the negation of (IPL) is true, and we've arrived at a contradiction. So (IPL) is not true. But this piece of reasoning constitutes an informal refutation of (IPL), and so (IPL) is informally refutable and we've proven it is by giving the argument, which means that we've proven (IPL). By informal reflection, (IPL) is true after all. Contradiction.

Another interesting example is *Informal Provability Curry* (IPC) sentence: a sentence saying of itself that if it's informally provable, then an arbitrary sentence φ is true (so, one gets different IPC sentences by substituting different sentences for φ):[7]

$$\text{If (IPC) is informally provable, then } \varphi. \tag{IPC}$$

Suppose (IPC) is false. Then the antecedent is true and (IPC) is informally provable, and, by informal reflection, true. So we informally proved (IPC), and thus have both the antecedent and the implication. Hence φ.[8]

6.2.2 Their formal counterparts and their use

Suppose we are working in a certain axiomatic theory $\mathtt{T}$, extending a sufficiently strong arithmetic. Relative to a fixed Gödel coding, by $\ulcorner\varphi\urcorner$ we mean the code of a formula φ. $Proof(x, y)$ is read as *x is a code of a sequence of codes of formulas which together constitute a $\mathtt{T}$-proof of a formula whose code is y*. $Proof(x, y)$ represents formal provability in $\mathtt{T}$, so that if indeed $Proof(m, n)$, $\mathtt{T} \vdash Proof(\bar{m}, \bar{n})$, and if $\neg Proof(m, n)$, $\mathtt{T} \vdash \neg Proof(\bar{m}, \bar{n})$ for any particular numbers m, n and their standard numerals $\bar{m}, \bar{n}$.

By $\mathtt{Pr}(y)$ we mean the formula $\exists x\, Proof(x, y)$, We will call it *the formal provability predicate of* $\mathtt{T}$. Recall that if $Proof(x, y)$ is constructed in a standard way,[9] this predicate has the following properties, usually referred to as *Hilbert-Bernays derivability conditions*:

$$\mathtt{T} \vdash \phi \Rightarrow \mathtt{T} \vdash \mathtt{Pr}(\ulcorner\varphi\urcorner) \tag{HB1}$$

$$\mathtt{T} \vdash \mathtt{Pr}(\ulcorner\phi \to \psi\urcorner) \to (\mathtt{Pr}(\ulcorner\phi\urcorner) \to \mathtt{Pr}(\ulcorner\psi\urcorner)) \tag{HB2}$$

$$\mathtt{T} \vdash \mathtt{Pr}(\ulcorner\phi\urcorner) \to \mathtt{Pr}(\ulcorner\mathtt{Pr}(\ulcorner\phi\urcorner)\urcorner) \tag{HB3}$$

A crucial theorem needed to prove the existence of the formalized counterparts of the paradoxical sentences is the Diagonal Lemma.

Theorem 91 (Diagonal Lemma)**.** *Let $\mathbb{N}$ be the standard model of natural numbers, then for every formula $\varphi(x)$ there is a sentence λ such that $\mathbb{N} \models \varphi(\ulcorner\lambda\urcorner) \equiv \lambda$.*

[7] Its construction is, obviously, inspired by the paradoxical Curry sentence which says of itself that if it is *true*, then so is an arbitrary sentence φ.

[8] Note that if you prefer a formulation in which we have 'φ is informally provable' in the consequent, it follows from (IPC).

[9] Readers interested in how provability predicates can be defined in the standard and non-standard ways can turn to (Smith, 2007; Halbach and Visser, 2014).

Proof. Let $Diag(x, y)$ be a formula representing the diagonal function,[10] Take an arbitrary formula $\varphi(x)$, define $\psi(y) := \forall z\,(Diag(y, z) \to \varphi(z))$ Let λ be the diagonalization of ψ, so that $\lambda \equiv \psi(\ulcorner\psi\urcorner)$. By the definition of diagonalization,

$$\lambda \equiv \forall z\,(Diag(\ulcorner\psi\urcorner, z) \to \varphi(z))$$

follows. But we know that the right-hand side of the equivalence is true iff z (the code of the diagonalization of ψ) has property φ, so $\lambda \equiv \psi(\ulcorner\lambda\urcorner)$. □

Diagonal Lemma can be strengthened to apply the provability in any sufficiently strong arithmetical theory $\mathtt{T}$, so that $\lambda \equiv \psi(\ulcorner\lambda\urcorner)$ is not only true in the standard model, but also provable in $\mathtt{T}$.

In what follows, instead of a standard provability predicate $\mathtt{Pr}$ we will often use a more generic P if we want to emphasize that a claim holds for *any predicate* satisfying certain conditions that $\mathtt{Pr}$ satisfies.[11] Later on, we'll use $\mathtt{B}$ as a specific provability operator introduced in our non-classical logic. We hope no confusion will arise in what follows.

Let's start with obtaining the formal counterparts of the problematic sentences discussed in Subsection 6.2.1. By Diagonal Lemma there are sentences γ, λ and ζ of the arithmetical language such that:

$$\mathtt{T} \vdash \gamma \equiv \neg\mathtt{Pr}(\ulcorner\gamma\urcorner) \tag{FPG}$$

$$\mathtt{T} \vdash \lambda \equiv \mathtt{Pr}(\ulcorner\neg\lambda\urcorner) \tag{FPL}$$

$$\mathtt{T} \vdash \zeta \equiv (\mathtt{Pr}(\ulcorner\zeta\urcorner) \to \varphi) \tag{FPC}$$

Now, (FPG) and (FPL) (or counterparts thereof for a given candidate for a provability predicate) can both be used to prove the following theorems that hold for any sufficiently strong arithmetical theory $\mathtt{T}$. The proofs are well known — we just go over them because we'll be making a point about them further on.

Theorem 92 (Montague)**.** $\mathtt{T}$*, if consistent, cannot contain (or be consistently extended to contain) a (possibly complex) predicate for which* (HB1) *and all instances of the reflection schema hold.*

Proof with (FPG). Suppose that there is such a predicate, call it P. Argue inside $\mathtt{T}$ using natural deduction:

1. $\gamma \equiv \neg P(\ulcorner\gamma\urcorner)$	Diagonal lemma
1.1 $\neg\gamma$	hypothesis
1.2 $P(\ulcorner\gamma\urcorner)$	equivalence elimination: 1,1.1
1.3 γ	Reflection schema: 1.2
2. γ	hypothesis discharge: 1.1 → 1.3
3. $P(\ulcorner\gamma\urcorner)$	(HB1): 2
4. $\neg P(\ulcorner\gamma\urcorner)$	equivalence elimination 1, 2
5. contradiction	3, 4

□

Proof with (FPL).

[10] It is the function which takes the code of a certain formula and returns the code of its diagonalization, so that if $n = \ulcorner\varphi(z)\urcorner$, then $diag(n) = \ulcorner\varphi(\bar{n})\urcorner$.

[11] So, for instance, when we say that P satisfies (HB1) we mean $\mathtt{T} \vdash \phi \Rightarrow \mathtt{T} \vdash P(\ulcorner\varphi\urcorner)$, etc.

1. $\lambda \equiv P(\ulcorner\neg\lambda\urcorner)$	Diagonal lemma
1.1 λ	hypothesis
1.2 $P(\ulcorner\neg\lambda\urcorner)$	equivalence elimination: 1,1.1
1.3 $\neg\lambda$	Reflection schema: 1.2
2. $\neg\lambda$	hypothesis discharge $1.1 \to 1.3$
3. $P(\ulcorner\neg\lambda\urcorner)$	(HB1): 2
4. $\neg P(\ulcorner\neg\lambda\urcorner)$	1, 2
5. contradiction	3, 4

□

Theorem 93 (Dual Montague)**.** $\mathtt{T}$*, if consistent, cannot contain (or be consistently extended to contain) a (possibly complex) predicate for which all instances of* $\varphi \to P(\ulcorner\varphi\urcorner)$ *(Provabilitation), and the Co-necessitation rule (if* $\mathtt{T} \vdash P(\ulcorner\varphi\urcorner)$*, then* $\mathtt{T} \vdash \varphi$*) hold.*

Proof with (FPG). Suppose that there is such a predicate, call it P. We use a natural deduction system. Argue inside $\mathtt{T}$:

1. $\gamma \equiv \neg P(\ulcorner\gamma\urcorner)$	Diagonal Lemma
1.1 γ	assumption
1.2 $\neg P(\ulcorner\gamma\urcorner)$	equivalence elimination: 1, 1.1
1.3 $\gamma \to P(\ulcorner\gamma\urcorner)$	Provabilitation for γ
1.4 $\neg\gamma$	MTT: 1.3, 1.2
2. $\neg\gamma$	assumption discharge: $1.1 \to 1.4$
3. $P(\ulcorner\gamma\urcorner)$	equivalence elimination: 1,2
4. γ	Co-necessitation: 3
5. contradiction	2,4

□

Proof with (FPL).

1. $\lambda \equiv P(\ulcorner\neg\lambda\urcorner)$	Diagonal Lemma
1.1 $\neg\lambda$	assumption
1.2 $\neg P(\ulcorner\neg\lambda\urcorner)$	equivalence elimination: 1,1.1
1.3 $P(\ulcorner\neg\lambda\urcorner)$	Provabilitation: 1.1
1.4 $\bot$	1.2, 1.3
2. λ	reductio ad absurdum: $1.1 \to 1.4$
3. $P(\ulcorner\neg\lambda\urcorner)$	equivalence elimination: 1, 2
4. $\neg\lambda$	Co-necessitation: 3
5. contradiction	2, 4

□

(FPC), the formal counterpart of (IPC), leads to Löb's Theorem:

Theorem 94 (Löb's Theorem)**.** *Let* $\mathtt{T}$ *be a sufficiently strong arithmetical theory containing a provability predicate* P *satisfying* (HB1)*,* (HB2) *and* (HB3)*. If* $\mathtt{T} \vdash P(\ulcorner\varphi\urcorner) \to \varphi$ *then* $\mathtt{T} \vdash \varphi$*.*

Proof. Suppose $\mathtt{T} \vdash P(\ulcorner\varphi\urcorner) \to \varphi$ and argue inside $\mathtt{T}$:

1. $\zeta \equiv (P(\ulcorner\zeta\urcorner) \to \varphi)$	Diagonal Lemma for $P(\ulcorner x\urcorner) \to \varphi$
2. $\zeta \to (P(\ulcorner\zeta\urcorner) \to \varphi)$	equiv elimination: 1
3. $P(\ulcorner\zeta \to (P(\ulcorner\zeta\urcorner) \to \varphi)\urcorner)$	(HB3): 2
4. $P(\ulcorner\zeta\urcorner) \to P(\ulcorner P(\ulcorner\zeta\urcorner) \to \varphi\urcorner)$	(HB2): 3
5. $P(\ulcorner\zeta\urcorner) \to (P(\ulcorner P(\ulcorner\zeta\urcorner)\urcorner) \to P(\ulcorner\varphi\urcorner))$	HB2 to the consequent: 4
6. $P(\ulcorner\zeta\urcorner) \to P(\ulcorner P(\ulcorner\zeta\urcorner)\urcorner)$	(HB3) for $P(\ulcorner\zeta\urcorner)$
7. $P(\ulcorner\zeta\urcorner) \to P(\ulcorner\varphi\urcorner)$	logic: 5,6
8. $P(\ulcorner\zeta\urcorner) \to \varphi$	proof assumption, 7
9. ζ	equiv elimination: 1, 8
10. $P(\ulcorner\zeta\urcorner)$	(HB1): 9
11. φ	8, 10

□

Notice, however, that these formal results, if one wants to take them as a guide to our understanding of informal provability, are rather disturbing. Reflection and (HB1) seem like plausible principles of informal provability, and yet, Theorem 92 suggests they cannot be consistently conjoined. Theorem 93 is not too scary, because we don't have strong intuitions supporting provabilitation. Löb's theorem, however, is worrying again. After all, we are inclined to think that *any mathematical claim whatsoever* is true, if provable, whereas Löb's theorem (well, the informal counterpart thereof, to be more precise) seems to suggest that this can hold only for those claims that are already provable.

6.3 On an informal reading of Löb's theorem

The philosophical aspects of Löb's theorem don't get discussed too often. The *locus classicus* is (Boolos, 1993, 54-55), where some reasons to be surprised by the theorem are discussed. Most of them are bad reasons: but just because Boolos discusses and criticizes bad reasons to be surprised with the theorem, it doesn't mean there aren't good reasons too. We'll quote Boolos *in extenso*, enumerating his points for further reference, and comment on them.

> Löb's theorem is considered utterly astonishing for at least five reasons:
>
> **Reason 1** "In the first place, it is often hard to understand how vast the mathematical gap is between truth and provability. And to one who lacks that understanding and does not distinguish between truth and provability, $\mathtt{Pr}(\ulcorner S\urcorner) \to S$, which the hypothesis of Löb's theorem asserts to be provable, might appear to be trivially true in *all* cases, whether S is true or false, provable or unprovable. But if S is false, S had better not be provable. Thus it would seem that S ought not always to be provable provided merely that (the possibly trivial-seeming) $\mathtt{Pr}(\ulcorner S\urcorner) \to S$ is provable."
>
> **Reason 2** "Secondly, `Pr` seems here to be working like negation. After all, if $\neg S \to S$ is provable, then so is S; proving S by

> proving $\neg S \to S$ is called *reductio ad absurdum* [...] Moreover, inferring S solely on the ground that $S \to S$ is demonstrable is known as begging the question, or reasoning in a circle. To one who conflates truth and provability, it may then seem that Löb's theorem asserts that begging the question is an admissible form of reasoning in **PA**."

Reason 3 "Thirdly, one might have thought that *at least on occasion*, **PA** would claim to be sound with regard to an unprovable sentence S, i.e., claim that *if* it proves S, then S holds. But Löb's theorem tells us that it never does so: **PA** makes the clam $\mathtt{Pr}(\ulcorner S \urcorner) \to S$ that it is sound with regard to S only when it obviously must, when the consequent is actually provable. As Rohit Parikh once put it, "**PA** couldn't be more modest about its own veracity"."

Reason 4 "Fourthly, one might very naturally suppose that provability is a kind of necessity, and therefore, just as $\Box(\Box p \to p)$ always expresses a truth if the box is interpreted as "it is necessary that" — for then $\Box(\Box p \to p)$ says that it is necessarily true that if a statement is necessarily true, it is true — $\mathtt{Pr}(\ulcorner(\mathtt{Pr}(\ulcorner S \urcorner) \to S)\urcorner)$ would also always be true or at least true in some cases in which S is false and not true only in the rather exceptional cases in which S is actually provable."

Reason 5 "Finally, it seems wholly bizarre that the statement that if S is provable, then S is true is not itself provable in general. For isn't it perfectly obvious, for any S, that S is true if provable? Why are we bothering with **PA** if its theorems are false? And how could any such (apparently) obvious truth not be provable?"

In **Reason 1** Boolos suggests that reflection might seem trivially true in all cases to someone who conflates truth and provability. **Reason 2** gives us a hint as to what Boolos had in mind. Apparently, confusing truth with provability (and thus $\mathtt{Pr}(\ulcorner S \urcorner)$ with S) makes one identify the inference legitimated by Löb's theorem:

$$\vdash \mathtt{Pr}(\ulcorner S \urcorner) \to S \;\Rightarrow \vdash S$$

with a fairly obviously incorrect inference pattern:

$$\vdash S \to S \;\Rightarrow \vdash S.$$

Thus, Boolos's critic of Löb's theorem is willing to reject the former in virtue of rejecting the latter. The fact that the conflation of truth and provability is represented as intersubstitutability of $\mathtt{Pr}(\ulcorner S \urcorner)$ and S suggests that the conflation meant in **Reason 1** consists in thinking that $\mathtt{Pr}(\ulcorner S \urcorner) \to S$ should hold simply because $S \to S$ does.

But this, we submit, is a straw-man position. No one who seriously defends the universality of reflection (that is, that it holds for all sentences) argues that $\mathtt{Pr}(\ulcorner S \urcorner) \to S$ simply because $S \to S$ holds. Rather, informal reflection seems compelling because the proof methods used in informal mathematics are

extremely reliable, so that if a mathematical claim is informally proven, this is considered sufficient evidence to accept that claim as true.

We don't have much to say about **Reason 3**, which we agree with — after all, it's a correct claim about *formal* provability contrasted with an intuition that we have about informal provability.

The argument considered in **Reason 4**, however, isn't too convincing and seems to be a straw-man argument for reflection. After all, to seriously argue against Löb's theorem holding for informal provability by saying that `Pr` is a kind of necessity, one would have to explain what it means *to be a kind of necessity*, why `Pr` indeed is one, and how it being a kind of necessity requires the relevant principle to hold.

In **Reason 5** Boolos briefly describes better grounds for accepting the universality of reflection. Alas, Boolos only brings up this dissonance and leaves it at that.[12] Nowhere in the book does he later come back to the topic to explain how this intuition about provability is to be squared with the harsh truth unveiled by Löb's theorem.

The reasons to be surprised by Löb's theorem according to Boolos aren't, surprisingly, the reasons we'd normally have to accept reflection. Let's try to explicate these grounds in more detail. What would happen if reflection was false? This would mean that some mathematical claim is informally provable and false. Any sane mathematician would say that this hypothetical situation is not possible since informal proofs are our only means of establishing the truth value of a mathematical claim: if an informal proof of a given sentence is correct, the sentence itself should be considered true.

Another argument for reflection goes as follow. Any (direct) informal proof can be divided into steps. Each step is either an instance of an axiom or a sentence that is already accepted as true, or follows from previous steps. The connection between any two steps in an informal proof is often expressed by phrases such as: "since φ holds, ψ is true" or "from φ it follows that ψ", or "it is easy to see that since φ is true ψ must be true". Clearly, these expressions say something about the relation between the truth values of φ and ψ, namely that from the fact that φ is true, we infer that so is ψ. It means that truth is preserved between any two steps in an informal proof. Hence, an informal proof as a whole has to preserve truth.[13]

So, in general, while there might be bad reasons to be surprised by Löb's theorem, and Boolos is right in criticizing them, there still seem to be good reasons to think reflection holds (and therefore, that Löb's theorem doesn't apply to informal provability).[14] The question now is: won't our commitment to reflection lead us into trouble? After all, it seems, when we accept reflection for informal

[12] Of course, giving a proof of Löb's theorem (which is what Boolos does) is a way of answering the question how such an obvious truth could fail to be provable. But such an answer pertains to formal provability only — the proof of Löb's theorem sufficiently explains why reflection fails for *formal* provability of, say, **PA**. The proof, however, doesn't cast much light on why we're still convinced that reflection is an obvious universal principle for *informal* provability, and how to square these two facts.

[13] When it comes to indirect proofs, without much effort one can easily translate any indirect proof into a direct one so the existence of indirect proofs does not cause any problems. A brute force technique is contraposition, but there are more elegant solutions as well.

[14] Or, in other words, that the reasoning justifying Löb's theorem for formal provability can't

provability, we end up with a paradox! Let's take a closer look at the best shot in this direction, fired by Graham Priest and JC Beall.

6.4 Informal provability meets paradoxes

6.4.1 The dialetheist argument formalized

Priest (2006, 46) and Beall (1999, 324) used (IPG) to argue that informal mathematics is inconsistent (and therefore, that a paraconsistent logic is needed).[15] This is, essentially, the argument that we've already discussed in subsection 6.2.1.

In this section, we'd like to zoom in on the argument and take a closer look by formalizing it.[16] Read $\Box$ as *it is informally provable that* and let $\vdash_I$ stand *in the metatheory* for *provability in informal mathematics*. The argument starts with the following rules for informal provability:

$$\vdash_I \Box\varphi \to \varphi \qquad \text{(Reflection)}$$

$$\text{If } \vdash_I \varphi, \text{ then } \vdash_I \Box\varphi. \qquad \text{(Nec)}$$

Note that $\vdash_I$ stands *in the meta-language* for being informally provable, whereas $\Box$ expresses informal provability *in the object language*. As for the first rule, according to Priest, it is analytic that if something is informally provable, it is true. Since this claim itself is analytic, it is informally provable. For the second principle, the intuition is that if something is provable, then its proof constitutes an informal proof of its provability.

The formalized argument relies also on the presence of self-reference in informal mathematics, so that it is supposed to be informally provable that (IPG) is equivalent $\neg\Box$(IPG) (for brevity, we'll use γ to stand for (IPG); no unclarity should arise).[17]

---

be extrapolated to analogous correct reasoning about informal provability.

[15] "Consider the sentence 'This sentence is not provably true.' Suppose the sentence is false. Then it is provably true, and hence true. By *reductio* it is true. Moreover, we have just proved this. Hence it is provably true. And since it is true, it is not provably true. Contradiction" (Priest, 2006, 46).

"Consider the sentence:

$$\gamma \text{ is (informally) unprovable.} \qquad (\gamma)$$

If γ is false, then it is provably true, and so true. By *reductio* γ is true. Since we have just (informally) proved γ, γ is (informally) provable. But since true, γ is (informally) unprovable. Contradiction" (Beall, 1999, 324).

[16] The formalization is inspired by the formulation of the argument contained in the new material in the second edition of *In contradiction* (Priest, 2006, 238).

[17] In a similar vein we could run the argument using (IPL), as the second proof of Theorem 92 suggests.

1. $\vdash_I \gamma \equiv \neg\Box\gamma$	self-reference
2. $\vdash_I \gamma \to \neg\Box\gamma$	equiv elimination: 1
3. $\vdash_I \Box\gamma \to \neg\gamma$	contraposition: 2
4. $\vdash_I \Box\gamma \to \gamma$	reflection
5. $\vdash_I \neg\Box\gamma$	logic: 3,4
6. $\vdash_I \gamma$	equiv elimination: 1,5
7. $\vdash_I \Box\gamma$	Nec: 6
8. Contradiction	5,7

Thus, it seems, informal mathematics is inconsistent.

Both (Nec) and (Reflection) are plausible inference patterns for informal provability. The only remaining potentially suspicious move is that in line 1 (we'll refer to it as (Premise 1). Notice that the vdash there is essential — if we replace (Premise 1) with mere truth of $\gamma \equiv \neg\Box\gamma$, we'll only be able to prove that γ is independent. So why would it be *informally provable* that γ is equivalent to $\neg\Box\gamma$?

6.4.2 Straightforward arguments for the equivalence fail

One way of arguing that it is, is to say that the availability of self-reference is an important feature of informal mathematics, and that '$\gamma \equiv \neg\Box\gamma$' is a sentence formulated in the language of informal mathematics. The strategy is suggested by the following extremely brief defense:

> There seems to be little hope of denying that γ is indeed a sentence of our informal mathematics. (Beall, 1999, 324)

Of course, 'being a sentence of...' is ambiguous. Beall might have simply meant that $\gamma \equiv \neg\Box\gamma$ is informally provable, in which case, it's not an argument, but a mere restatement of the premise. On the other hand, he might have meant that since γ is formulated in the language of informal mathematics, and we somehow recognize $\gamma \equiv \neg\Box\gamma$ to be true, the equivalence, by the same token, is informally provable.

One way to react is to deny that sentences containing "informally provable" are sentences *of* informal mathematics. This strategy has been pursued by Tanswell (2016), who in his criticism of the dialetheist interpretation of Gödelian phenomena, denies the premise and insists that γ is not formulated in the language of informal mathematics:

> I take the concept of informal proof to be used to talk and reason about mathematics without it being a part of mathematics. Of course, I hold that informal proof and provability are very important notions in talking about mathematics, but it is crucial to emphasize that these are notions about mathematics. To establish that the paradox will render mathematics inconsistent, though, we need the extra claim that it is a part of informal mathematics. (Tanswell, 2016, 163)

His reasons for denying it are that (i) informal provability lacks a precise mathematical definition, and that (ii) it doesn't relate with other mathematical con-

cepts in the way standard mathematical concepts, such as that of a group or that of integer, do.

Our impression is that this strategy might be too sweeping. It excludes any claim involving the concept of informal provability from being informally provable, whereas in our uninspired moments we like to think that, say, claims such as 'any mathematical claim either is, or isn't informally provable' is as close to being informally provable as it gets.

We also aren't too convinced by Tanswell's reasons for his sweeping rejection. "Relating with other mathematical concepts the way standard mathematical concepts do" lacks a precise mathematical definition either, and it's rather unclear whether being unlike the concept of numbers or algebraic groups excludes a concept from the domain of informal mathematics. Consider computability theory: it also deals with concepts (algorithm, computation, ...) quite unlike that of algebraic groups or integers: does this mean it's not part of informal mathematics? And lacking a precise mathematical definition is also not a necessary feature of many concepts used in informal mathematics. For centuries, various concepts were used in mathematics, despite their precise definitions not being available. If we were to follow Tanswell's advice, many things we uncontroversially consider mathematical developments wouldn't deserve that name (think for instance about the development of the currently standard notions of sets, limits, functions, and so on).

Perhaps, indeed, we have no good grounds to deny that self-reference is available in informal mathematics, in the sense that $\gamma \equiv \neg\Box\gamma$ indeed is formulated in the language of informal mathematics. After all, we are (at least *prima facie*) free to label sentences as we wish, and free to formulate sentences containing such labels. But this on its own doesn't entail that the equivalence is *provable* in informal mathematics. There is a more involved sense in which self-reference is required to be available in informal mathematics for (Premise 1) to hold: the sense in which such availability would make the equivalence *provable*. But what reasons do we have to think that self-reference in this sense indeed is available in informal mathematics?

The dialetheist cannot leave (Premise 1) undefended, and vague reference to 'self-reference being available' is not enough. What more can they do? Let's see. One way to argue is that in light of Tanswell's strategy being too sweeping, there is no reason to deny that (Premise 1) is formulated in the language of informal mathematics. Secondly, one could continue, the claim can be proven the way most informal mathematical claims are proven. It's simply enough to approach the blackboard and write appropriate stuff on it! I simply write:

$$(\gamma) \quad \neg\Box\gamma$$

and you immediately can recognize the equivalence used in (Premise 1) as true. I introduced γ *as* (equivalent to) the sentence $\neg\Box\gamma$, so by the power of this very move, γ is equivalent to $\neg\Box\gamma$. Q.E.D.!

This, however, is too hasty. Why would such an introduction have the power to establish a mathematical theorem? Is, perhaps, the claim supposed to be true by definition? There is no good reason to think that all ways of "introducing" new symbols automatically translate into informal provability in mathematics

(think for instance about tonk and plonk etc. (Belnap, 1962)). Quite crucially, observe that the introduction of (γ) doesn't yield a proper definition, in the sense that when a proper definition of a new propositional constant σ is introduced to extend a language $\mathcal{L}$, it has to be introduced as equivalent to a sentence *of* $\mathcal{L}$ *not containing* σ, otherwise the definition would be circular. Once the introduction fails to have the format of a proper definition,[18] it's far from clear whether merely being written on a blackboard results in the equivalence being informally provable.[19]

Well, perhaps, one could insist that in some domains of informal mathematics some non-well-founded definitions live peacefully with unicorns and rainbows without causing any trouble (Aczel, 1988). Maybe. But in such contexts, showing that such a definition doesn't pose a threat actually requires some work, and any argument that would use the definition to obtain inconsistency would rather be taken as a sign of pathology of the definition, not of the contradiction of informal mathematics.

So maybe let's abandon the idea that we should think of the introduction of (γ) as a definition, and let's think of it only as a postulation, noticing that at least certain mathematical postulates can be true without being definitions. On this reading, even though it perhaps isn't true that (Premise 1) holds by definition, it still is supposed to be true because we postulated it to be so! But again, one needs to be more careful. Not just any postulation on a blackboard leads to a new informally provable statement. Otherwise, I could just approach the blackboard, write:

$$\text{Let } x = 0 \wedge x = 1,$$

and prove the inconsistency of mathematics from this postulate (and, say, Peano Arithmetic).

One might push further, though, by insisting: *but this isn't just any postulation! I displayed a formula,* $\neg\Box\gamma$*, and introduced an abbreviation for it,* γ*. Obviously, any abbreviation of a formula, by the very fact of being introduced, is equivalent to the formula being abbreviated.*

Yes. Mathematicians do use abbreviations in their work. But if you're serious about relying on the mathematicians' practice of using abbreviations, then perhaps you should also notice that sane working mathematicians don't use as abbreviations the sub-formulas of formulas being abbreviated (and in general, if a symbol is a meaningful part of the expression being abbreviated, it's not afterwards used as an abbreviation for the whole expression).

But even if we pretend that the act of introducing (γ) is an act of introducing an acceptable abbreviation, (Premise 1) still doesn't follow. Imagine a mathematician correctly introduces an expression, say τ, as an abbreviation for a longer formula, say $\Box\Box\Box\varphi \to \Box\Box\varphi$. Now ask them:

[18] Note also that often even a proper definition has to be accompanied by an existence and uniqueness proofs; this isn't too relevant in the case of propositional constants, though. We mention this only to emphasize that even the introduction of proper definitions isn't to be take too lightly in informal mathematics.

[19] Of course, mathematicians admit methods of introducing concepts which go beyond proper definitions. For instance, inductive definitions have a somewhat different form. But the reliability of such "definitions" piggybacks on previously proven theorems showing that in principle they can be replaced by proper ones.

Have you therefore informally proven that $\tau \equiv [\Box\Box\Box\varphi \to \Box\Box\varphi]$?

The reply would most likely be preceded (or followed) with a deep stare in your direction, and it would be somewhere close to: *no, I haven't proven anything, it's just an abbreviation I introduced for the sake of convenience.* So, it seems, simple strategies of defending (Premise 1), fail. Can the dialethist do better?

6.4.3 Who ya gonna call? Diagonal lemma!

At this point, a persistent dialetheist might conjure the diagonal lemma. One could first argue that informal mathematics is recursively axiomatizable, because finite human beings can be taught to do informal mathematical proofs:

> The naive notion of proof is a social one. In particular, it is one which is taught and, correspondingly, learned. Yet the collection of proofs is (potentially) infinite. Hence the notion cannot be taught by giving a simple finite list. If proof is not a recursive notion, then the process whereby it is learned becomes unintelligible. Consider the following analogy. People are able to produce (potentially) infinitely many numerals. Moreover, everyone can agree that what is produced is a numeral. This is perfectly understandable in virtue of the fact that numerals can be produced by applications of effective rules to a finite vocabulary. (They are a recursive class.) If this were not the case, then that agreement is achieved would be a mysterious and even mystical process. So it is with proof. (Priest, 2006, 41)

The argument continues: since `M` is a formal theory which satisfies the standard conditions required for the Diagonal Lemma, we can prove the Diagonal Lemma for `M`. Once we do so, the existence of γ satisfying (Premise 1) follows.

The move from the learnability of a concept to its recursive axiomatizability is somewhat hasty. After all we also seem to understand some notions that aren't recursively axiomatizable. For instance, we tend to think that we understand what it means for a first order arithmetical sentence to be true, even though the set of first order arithmetical truths isn't recursively axiomatizable.[20]

Perhaps there is a stronger sense in which we can learn to do mathematics, in which we can't learn the concept of first order mathematical truth, a sense in which it follows that informal mathematics is recursively axiomatizable? What would it be? For one thing, saying that for any given claim a mathematically competent human (or a group thereof) can decide whether it's informally provable would be too much. There are many open problems in mathematics where it is not known if a claim is provable, and it is not even clear whether it will ever be known.

Well, maybe the difference is that the notion of informally provable claims is "semi-decidable", in the sense that if a claim is provable, it would be eventually recognized as such by means of some general procedure? If informal mathematics was recursively axiomatizable, such a procedure would exist: simply start

[20] See (Tanswell, 2016) for criticism from another angle.

effectively listing all proofs, and if a claim is provable, you'll run into its proof eventually.

But even this feat seems out of reach of a mathematically competent human being or a group thereof: notoriously, some mathematical proofs have eluded generations of mathematicians, and their invention wasn't a matter of following some general recipe. Equally notoriously, we're as far from inventing a *general* procedure for proving any informally provable claim as we were thousands years ago.

So there seems to be no recipe for deciding whether a claim is informally provable, and not even a recipe for finding an informal proof of a claim if it is possible. In what other sense can we connect the "learnability" of mathematics with recursive axiomatizability? One intuition for the recursive axiomatizability stems from the observation that informal proofs can be checked for correctness in a finite amount of time:

> Consider the situation which arises if the notion of proof is non-effective. There is then no certain means by which, when a sequence of formulas has been put forward as a proof, the auditor may determine whether it is in fact a proof (Church, 1996, 53).

In the context of *formalized* theories, the decidability of the set of proofs entails the semi-decidability of theories: to identify a provable claim as such, *start listing all potential proofs* checking them for correctness as you go, and eventually you'll find the proof of the theorem that you're after.

The question is, however, whether the notion of proof that is supposed to be effective in fact is the notion of informal provability in mathematics in general. Yes, indeed, at any particular stage of development of mathematics, proofs (and flawed alleged proofs) given by mathematicians are supposed to be recognizable as such (even though, sometimes not without effort and not by a single lone mathematician). This, however, doesn't mean that at any particular stage of development of mathematics one could simply sit down and list all admissible ways of proving mathematical theorems, independently of the development of mathematics.[21] Ask yourself how good a job would Euclid do if asked to list all means of proving things in mathematics (hint: not an awesome one). Now, why should we think that we're now in a better position to predict the development of mathematics? In other words, even if we can decide whether a given supposed proof is a proof (and even this is a strong idealization), we can't start listing all potential proofs, because there are ways of proving things that we can't predict, because they depend on creative developments of the mathematician's toolbox, or novel connections between different fields of mathematics etc., and there is no recipe for listing these *a priori*.

Informal mathematics evolves: at any stage we have, perhaps, a decent grasp of what would count as a correct proof; but new domains, new techniques, new

[21] The phenomenon is quite general: recursive enumerability is not generally closed under countable unions. For example, consider partial truth predicates for restricted formula complexity, each of which is definable in the arithmetical language, while the union of their extensions isn't. For another example, consider the fact that Peano arithmetic proves the consistency of each of its finite sub-theories, but it does not prove the consistency of their union. For a thorough study of related phenomena see (Franzén, 2004).

methods, new axioms, new definitions, new re-conceptualizations etc. can always be introduced in a fashion that doesn't seem to be algorithmic.[22] Such developments are mirrored by an appropriate extension or modification of the notion of a proof, and it's at least far from obvious that such developments can be at any particular stage captured by a recursive axiomatization.

Perhaps, the claim is simply obvious? Presumably, it would go along the following lines:

> For any (mathematical) property there is a mathematical sentence which is provably (in informal mathematics) equivalent to the claim that it has the property.

Does this sound plausible? Not really — buying into supposedly intuitive claims involving quantification over all properties has been passé for at least a century, and the fashion doesn't seem to be coming back anytime soon.

One might insist that the above reading is too rough. After all, the original lemma is about codes of formulas and about arithmetical properties. So perhaps the official informal counterpart of the Diagonal Lemma, instead of mentioning properties and attributing them to claims, should rather talk of *codes* of mathematical formulas and their *arithmetical* properties. In this setting, we no longer have to worry about the notion of property being suspiciously wide, and about mathematical properties being attributable to sentences. Also, the proof of the Diagonal Lemma isn't too complicated, and, one could argue, once we think about informal mathematics in terms of codings, we should be able to repeat its steps, *mutatis mutandis*, for informal provability as well.

But would we? The original proof relies heavily on the availability of coding of the language and inferential steps of a formal mathematical theory. Can we effectively code the language and all potential inferential moves of informal mathematics? Not really: the language and methods of informal mathematics are indefinitely extensible and we can't predict what devices and expressions it will comprise in, say, two centuries, just as Euclid couldn't have predicted the development of category theory or computer-assisted proofs.

But wait! We suggested that at least, *currently used* methods of proof might be claimed to be recursively axiomatizable in virtue of us, finite beings, being (supposedly) able to evaluate any mathematical proof formulated using currently available methods that life can throw at us.[23] Can't the dialetheist run with that assumption?

Well, not really. Say you read $\Box$ and $\vdash_I$ as expressing informal provability by means of *currently* used mathematical methods. How would you go about defending (Premise 1)? The straightforward strategies discussed in section 6.4.2 still fail for exactly the same reasons.

But what about the Diagonal Lemma? Even if we admit that the strategies used in the proof of the Diagonal Lemma for a formal system of arithmetic

[22] Well, we can't *prove* the class of all such methods really isn't algorithmic, because we'd have to pin their set down first, and it's exactly our point that there is no known way of going about this.

[23] However, for an explanation why such a formalization is not forthcoming anytime soon, and why there wouldn't be a unique one anyway, see (Tanswell, 2016).

belong to the current mathematical toolbox, it still doesn't follow that we can run an analogous proof for the whole of today's mathematics and still end up with something that's provable by means of today's mathematics. The problem is, running such an analogous proof properly would first require actually writing down the axiomatic system capturing the whole of today's mathematics — and doing *that* certainly isn't a standard proving technique of today's mathematics. If you buy into this variant of the dialetheist argument, you might equally well embrace dialetheism because of Richard's paradox.

So, we conclude, the strategy of defending (Premise 1) by conjuring some variant of the Diagonal Lemma that would apply to informal mathematics, fails, because informal mathematics is too lively a beast to be tamed.

Where does this leave us? The dialetheist argument for the inconsistency of informal mathematics fails. The formalizations of the supposedly paradoxical claims, however, yield serious and interesting limitative results concerning formalized theories. We don't have a sufficient reason to think that we can capture all of (future, past and present) informal mathematics by means of a formal axiomatic system. Are we done?

Not quite. While we might be unable to axiomatize informal mathematics, we still might ask what inferential principles hold for informal provability and try to formulate a formal axiomatized theory capturing those principles. This is the task that we'll turn to now.

6.5 A non-deterministic logic of provability

6.5.1 Motivations

Perhaps we convinced the reader that the notion of informal provability isn't as paradox-ridden as the dialetheist would like it to be. Still, however, both the discussion in Section 6.4 and the results of Section 6.2 might suggest severe pessimism regarding any possibility of a formal grasp of the properties of informal provability.

We should not despair (or at least not too much), though. The discussion of Section 6.4 indicates only that we shouldn't hope for a recursive axiomatization of all informally provable mathematical sentences. This doesn't preclude the possibility of developing a logic that captures *formal* properties and *valid inference principles* of the informal provability operator. We've already seen, for instance, that reflection is a principle that seems plausible in this context.

Historically, the first candidate for a provability logic, has already been suggested as such by Gödel. It was the modal logic **S4**, which contains as axioms all the substitutions of classical tautologies in the language with $\Box$, all substitutions of the schemata:

$$\begin{array}{ll} \text{(K)} & \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi) \\ \text{(M)} & \Box\varphi \to \varphi \\ \text{(4)} & \Box\varphi \to \Box\Box\varphi \end{array}$$

and is closed under two rules of inference: *modus ponens* (from $\vdash \varphi$ and $\vdash \varphi \to \psi$ infer $\vdash \psi$), and *necessitation* (Nec): if $\vdash \varphi$, then $\vdash \Box\varphi$.

The principles of **S4** seem sensible when $\Box\varphi$ is read as 'it is provable that φ': if an implication and its antecedent are provable, then so is its consequent, whatever is provable should be true, and if something is provable, we can prove that it is by simply displaying the proof. The system was used in 1933 by Gödel to interpret intuitionistic propositional calculus (which is closely related to reasoning about provability).

Prima facie **S4** is a viable candidate for the logic of informal provability. All its principles are intuitive as principles of informal provability.[24] Yet, **S4** falls pray to Montague's theorem. This means that we cannot consistently have a provability *predicate* for which all the properties captured by the **S4** axioms hold, as long as we have sufficiently rich arithmetic in the background.

On the other hand, avoiding the difficulty by staying at the level of informal provability *sentential operator*, as long as the strategy can't be extrapolated to the first order level, seriously limits expressive power of our theory of informal provability.

The question now is: does this mean that the enterprise of developing a logic of informal provability which wouldn't, so to speak, get aggressive when put in one cage with an arithmetical theory, is doomed? Moreover: can we develop such a logic motivated not only by the negative desire to avoid contradiction, but also by positive and sensible independent intuitions? The answer, we think, is positive.

6.5.2 The strategy

Philosophers faced a similar problem when constructing a formal theory of truth. There is an intuitive schema governing the truth-predicate: the **T-schema**

$$T(\ulcorner\varphi\urcorner) \Leftrightarrow \varphi,$$

where T is a truth-predicate. It is well-known, by Tarski's undefinability theorem, that not all instances of this schema can be consistently added to Peano arithmetic.

One of the most common solutions to this problem is to weaken the background logic to a non-classical one. By choosing an interesting logic, it's possible to circumvent the problem and to have a decent formal theory incorporating an interesting class of instances of the T-schema. If only we supplement this picture with a convincing and independently motivated philosophical story explaining why this particular non-classical logic should be used, our job as philosophical logicians is done. We will follow a similar route. Most notably, our goal is to explore the option of treating mathematical provability as a partial notion, just as some prominent theories of truth treat the truth predicate as a partial one. After all, there is an intuitive division of mathematical claims into provable, refutable and undecidable.

In the standard Kripke construction for truth one relies on the Strong Kleene logic to deal with the partial truth predicate. Alas, Strong Kleene Logic is

[24] Whether **S4** is a complete system of informal provability is still an open problem; see (Leitgeb, 2009) for a more elaborate discussion.

not appropriate for modeling informal provability, for it seems that informal provability is not truth-functional. For instance, it is sometimes, but not always the case that a disjunction of two sentences independent of a given theory is independent of that theory. For this reason, it seems that no deterministic many-valued logic could do the job. We have to do something else then.

Let $\mathcal{L}$ be a propositional language (understood as the set of all well-formed formulas) constructed from propositional variables $Var = \{p_1, p_2, \dots\}$ and Boolean connectives $(\neg, \wedge, \vee, \rightarrow, \equiv)$ in the standard manner. We will use Greek letters $\varphi, \psi, ...$ as meta-variables for formulas. The language that results from extending the set of Boolean connectives with one unary operator $\mathtt{B}$ will be denoted $\mathcal{L}_{\mathtt{B}}$. We will use $\mathtt{B}$ to express provability within the object language.

By an *assignment* we mean any function $v : Var \mapsto Val$, where Val is a set of values. By an *evaluation* e_v built over an assignment v we will mean a function assigning values to all well-formed formulas $(e_v : \mathcal{L}_{\mathtt{B}} \mapsto Val)$ agreeing with v on Var (propositional variables), and satisfying some additional constraints determined by a given logic.

In the case of standard classical propositional logic, evaluations are unambiguously determined by assignments. For each assignment there is exactly one evaluation extending it.

It is possible to construct sensible logics for which this uniqueness fails. One nice example is the propositional variant of paraconsistent logic CLuN (Batens and De Clercq, 2004).[25] The standard semantics of CLuN is similar to the semantics of classical propositional logic with one difference: the truth conditions for negation are different.

Both for classical logic and for CLuN we have $Val = \{0, 1\}$. In classical propositional logic $e_v(\neg\varphi) = 1$ iff $e_v(\varphi) = 0$. In CLuN this equivalence is weakened to an implication: if $e_v(\varphi) = 0$, then $e_v(\neg\varphi) = 1$. (Clauses for the rest of connectives are the same as in classical propositional logic.) In other words, CLuN allows for gluts for negation: both φ and $\neg\varphi$ can be true in one and the same evaluation.

The standard semantics of CLuN has another interesting feature. It is non-deterministic: assignments of values to propositional variables do not uniquely determine evaluations of all formulas. One and the same assignment might be extended in different ways to different evaluations, as long as they obey classical clauses for connectives other than negation and the implication above for negation. For instance, if $v(p) = 1$, there is one evaluation e^1_v such that $e^1_v(\neg p) = 0$ and there is another one e^2_v such that $e^2_v(\neg p) = 1$.

6.5.3 Non-deterministic semantics

We apply a similar trick to develop a non-deterministic semantics for a logic which would help us model the notion of informal provability. The logic will be three-valued: we take the set of values $Val = \{0, n, 1\}$. The intended interpretation of the values is as follows. 1 stands for *(informally) provable*, 0 represents *(informal)*

[25] A general framework for non-deterministic logics can be found in (Avron and Zamanski, 2011). Particular systems discussed there have, however, quite different motivation from ours, and quite different matrices.

refutability and n stands for *being neither (informally) provable, nor (informally) refutable.*

Here's the semantics for connectives of $\mathcal{L}_{\mathtt{B}}$ by means of non-deterministic matrices. Let's start with negation. A given formula is informally provable iff its negation is informally refutable. A given formula is informally refutable iff its negation is informally provable. A formula is undetermined iff its negation is.

For disjunction we introduce non-deterministic clauses. The equivalence:

$$e_v(\varphi \vee \psi) = 1 \text{ iff } e_v(\varphi) = 1 \text{ or } e_v(\psi) = 1$$

is weakened to one direction only:

$$\text{If } e_v(\varphi) = 1 \text{ or } e_v(\psi) = 1 \text{ then } e_v(\varphi \vee \psi) = 1.$$

The intention behind the introduction of non-determinism is this. We want to allow for the possibility of there being informally (absolutely) undecidable mathematical sentences (without saying that there are any). Yet, even for such sentences (if there are any), some disjunctions built from them might be informally undecidable, while some others will be informally provable. Say φ and ψ are informally undecidable (and therefore, so is $\neg\varphi$). Then, while we might think that $\varphi \vee \psi$ is informally undecidable, we might be inclined to think that $\varphi \vee \neg\varphi$ is informally provable despite φ not being decidable.

For instance, you might be inclined to think that the Continuum Hypothesis (CH) is informally undecidable, while $CH \vee \neg CH$ is still informally provable, being a logical truth. This however, clearly doesn't mean that $CH \vee CH$ is provable, and so not every disjunction of undecidable sentences is decided.

Once we gave semantics for negation and disjunction, matrices for other Boolean connectives follow. Conjunction $\varphi \wedge \psi$ is taken to have the same matrix as $\neg(\neg\varphi \vee \neg\psi)$.

The motivation for the non-deterministic case for conjunction is this. For some informally undecidable sentences we may be able to prove that they are mutually contradictory, which makes their conjunction informally refutable. For some others it may be impossible, and so their conjunction remains informally undecidable.[26] Implication is taken to have the same matrix as $(\neg\varphi \vee \psi)$, and equivalence has the same matrix as $((\varphi \to \psi) \wedge (\psi \to \varphi))$.

The intended reading of $\mathtt{B}\varphi$ is 'φ is informally provable.' The matrix for $\mathtt{B}$ is non-deterministic. The intuition behind this move is the following. If a formula is informally provable ($e_v(\varphi) = 1$), then giving its own proof is also a proof of its provability ($e_v(\mathtt{B}\varphi) = 1$), and the other way around. If a formula is informally refutable $e_v(\varphi) = 0$, then giving its own refutation is also a refutation of its provability ($e_v(\mathtt{B}\varphi) = 0$). If a formula is informally undecidable ($e_v(\varphi) = n$), then one of two things may happen. First, it may be the case that the undecidability of that formula is informally provable, and so its informal provability is refutable ($e_v(\mathtt{B}\varphi) = 0$). Second, it may be the case that its absolute informal undecidability

[26] Notice that just because $\varphi \wedge \psi$ has the same truth table as $\neg(\neg\varphi \vee \neg\psi)$, it doesn't follow that the substitution of expressions of this form preserves the value under an interpretation. This will fail due to indeterminacy. (The substitutability will be regained once we move from BAT logic to CABAT logic.)

is not informally provable, and so its absolute informal provability is informally undecidable ($e_v(\mathtt{B}\varphi) = n$).

All these conditions are captured by the following tables:

$\neg$	φ
0	1
n	n
1	0

$\vee$	0	n	1
0	0	n	1
n	n	n/1	1
1	1	1	1

$\wedge$	0	n	1
0	0	0	0
n	0	0/n	n
1	0	n	1

$\rightarrow$	0	n	1
0	1	1	1
n	n	n/1	1
1	0	n	1

$\equiv$	0	n	1
0	1	n	0
n	n	0/n/1	n
1	0	n	1

B	φ
1	1
n/0	n
0	0

Because we interpret value 1 as **B**eing an **A**bsolute **T**heorem (BAT), we call the logic thus obtained BAT-logic and we'll use the bat symbol 🦇 to denote its consequence relation, which we define as follows.

A BAT-assignment v is a function from propositional variables W to $\{0, n, 1\}$. A BAT-evaluation over an assignment v is a function which assigns values to all formulas of L, agrees with v on W and obeys the constraints we gave for the connectives. Notice that due to non-deterministic clauses, one and the same assignment might underlie multiple evaluation functions.

By Γ🦇φ, where Γ is a set of formulas, we will mean that any BAT-evaluation which assigns 1 to all formulas in Γ assigns 1 to formula φ. We say that φ is a BAT-tautology iff $\emptyset$🦇φ. We say that φ is a BAT-countertautology iff $\emptyset$🦇$\neg\varphi$.

Observe that disjunction is neither commutative nor associative. Take the assignment v where all propositional variables have value n and consider two formulas: $\varphi = p \vee q$ and $\psi = q \vee p$. As far as φ and ψ are concerned, there are four possible ways to extend this assignment:

$$e_v^1(\varphi) = n = e_v^1(\psi)$$
$$e_v^2(\varphi) = 1, e_v^2(\psi) = n$$
$$e_v^3(\varphi) = n, e_v^3(\psi) = 1$$
$$e_v^4(\varphi) = 1 = e_v^4(\psi).$$

BAT logic is too weak to eliminate extensions (e_v^1, e_v^2, e_v^3), in which φ and ψ obtain different values, and which show that neither φ🦇ψ, nor ψ🦇φ. Thus, it needs to be strengthened.

6.5.4 Strengthening BAT

Usually, to obtain a stronger logic from a logic with non-deterministic semantics we have to limit the range of available possible extensions of given assignments.[27] In our case, it will be particularly useful to enrich one logic (the non-deterministic one) by another one (classical logic), which will be used to "filter out" certain assignments.

[27]The most common way to strengthen a non-deterministic logic is to use the level-evaluation method (Coniglio et al., 2015) Due to simplicity, we prefer our method.

Definition 95. Let L be a logic. We say that a BAT-evaluation e belongs to the L-filtered set of BAT-evaluations just in case the following conditions hold:

1. For any two formulas φ, ψ, if $\models_L \varphi \equiv \psi$ then $e(\varphi) = e(\psi)$,
2. For any L-tautology φ, $e(\varphi) = 1$,
3. For any L-countertautology φ, $e(\varphi) = 0$.

We will focus on the case where L is classical logic (L=CL), and we simply use $\models$ to denote the classical consequence relation. By $\Gamma \vdash_{CL} \varphi$ we will mean that for any evaluation e in the CL-filtered set of BAT-evaluations if $e(\psi) = 1$ for all $\psi \in \Gamma$ then $e(\varphi) = 1$. The resulting logic is called CLBAT.

We may also be inclined to strengthen BAT-logic in a different manner. A quite intuitive way to go is to close BAT-logic under classical consequence. It can be done by the following condition:

Definition 96 (Closure condition)**.** An extension of BAT (in $\mathcal{L}_{\mathtt{B}}$) satisfies the closure condition just in case for all $\mathcal{L}_{\mathtt{B}}$-formulas $\varphi_1, \varphi_2, \cdots, \varphi_k, \psi$ such that

$$\varphi_1, \varphi_2, \cdots, \varphi_k \models \psi,$$

where $\models$ is the classical consequence relation for $L_{\mathtt{B}}$, for any BAT-evaluation e_v, if $e_v(\mathtt{B}\varphi_i) = 1$ for any $0 < i \leq k$, then $e_v(\mathtt{B}\psi) = 1$.

The result of closing BAT-logic under the closure condition will be called CABAT logic and its consequence relation will be denoted by $\vdash_C$. It turns out that the above conditions are equivalent, and so are the resulting logics:

Theorem 97. $\Gamma \vdash_C \varphi$ *iff* $\Gamma \vdash_{CL} \varphi$.

6.6 Basic Properties of CABAT

Quite trivially, CABAT logic is strictly stronger than BAT logic. The first interesting thing to see is that Deduction Theorem is not generally valid in CABAT:

Theorem 98. *If* $\vdash_C \varphi \to \psi$ *then* $\varphi \vdash_C \psi$ *but the implication in the opposite direction doesn't hold in general.*

In CABAT implications are stronger than the corresponding consequence relation, simply because the consequence relation informs us only about those evaluations in which all the premises have value 1. For instance, the consequence relation $\varphi \vdash_C \psi$ does not determine the value of implication $\varphi \to \psi$ when both ψ and φ have value n. On the other hand, $\vdash_C \varphi \to \psi$ uniquely determines the value of the implication under the previous assignment.

Lack of the deduction theorem makes a difference when we look at inference patterns with provability operator. Usually, principles for provability are valid in CABAT as consequence relations whereas their implicational formulations may be invalid. We are not terribly worried about that, since given our reading $\varphi \vdash_C \psi$ means that if φ is informally provable then φ is and this is the phrase

we intended to formalize. On the other hand, $\vdash_C \varphi \to \psi$ says: *if φ is informally provable, then so is ψ and if the antecedent is undecidable then the consequent is either provable or independent*, which is a stronger claim than $\varphi \vdash_C \psi$. So, keep in mind that implication no longer expresses the natural language conditional (as if it ever did), or the implication that we'd like to formalize in the system.

Let's take a look at some schemata involving the provability predicate and we will indicate whether they are sound for informal provability. The table below summarizes which inference patterns are valid in CABAT and whether the principle, according to us, is intuitive or not (we add question mark in cases where one might be, at least *prima facie*, worried):

Principle	Valid?	Intuitive?
$(\mathtt{B}\varphi \wedge \mathtt{B}\psi) \vdash_C \mathtt{B}(\varphi \wedge \psi)$	Yes	Yes
$\mathtt{B}(\varphi \wedge \psi) \vdash_C (\mathtt{B}\varphi \wedge \mathtt{B}\psi)$	Yes	Yes
$\mathtt{B}(\varphi \vee \psi) \vdash_C (\mathtt{B}\varphi \vee \mathtt{B}\psi)$	No	No
$(\mathtt{B}\varphi \vee \mathtt{B}\psi) \vdash_C \mathtt{B}(\varphi \vee \psi)$	No	?
$\varphi \vdash_C \mathtt{B}\varphi$	Yes	Yes
$\mathtt{B}\varphi \vdash_C \varphi$	Yes	Yes
$\mathtt{B}\varphi \vdash_C \neg\mathtt{B}\neg\varphi$	Yes	Yes
$\mathtt{B}\varphi \vdash_C \mathtt{BB}\varphi$	Yes	Yes
$\mathtt{BB}\varphi \vdash_C \mathtt{B}\varphi$	Yes	Yes
$\mathtt{B}(\varphi \to \psi) \vdash_C (\mathtt{B}\varphi \to \mathtt{B}\psi)$	No	?
$\mathtt{B}(\varphi \to \psi), \mathtt{B}\varphi \vdash_C \mathtt{B}\psi$	Yes	Yes
$\mathtt{B}(\varphi \wedge \neg\varphi) \vdash_C \mathtt{B}(\psi)$	Yes	Yes
$\mathtt{B}\varphi \vee \mathtt{B}\neg\varphi$	No	No
$\mathtt{B}\varphi \vee \neg\mathtt{B}\varphi$	Yes	Yes
$\neg\mathtt{B}\varphi \vdash_C \mathtt{B}(\neg\varphi)$	No	No
$\mathtt{B}(\neg\mathtt{B}\varphi) \vdash_C \mathtt{B}(\neg\varphi)$	No	No
$\mathtt{B}(\neg\mathtt{B}\neg\varphi) \vdash_C \neg\mathtt{B}(\neg\mathtt{B}\varphi)$	No	No

One thing that might seem worrying is that

$$(\mathtt{B}\varphi \vee \mathtt{B}\psi) \not\vdash_C \mathtt{B}(\varphi \vee \psi)$$

After all, if φ is informally provable, shouldn't $\varphi \vee \psi$ also be informally provable? This worry, however, stems from the fact that the provability of a disjunction in CABAT says something weaker than that one of its disjuncts is provable — after all, $\varphi \vee \neg\varphi$ is going to be informally provable without either φ or $\neg\varphi$ being informally provable. So, we submit, the intuition should be rather captured by requiring that the following should hold:

$$\mathtt{B}\varphi \vdash_C \mathtt{B}(\varphi \vee \psi)$$
$$\mathtt{B}\psi \vdash_C \mathtt{B}(\varphi \vee \psi)$$

and indeed, they do.

Another worry might be that the following asymmetry between at least *prima facie* close cousins can be observed:

$$\mathtt{B}(\varphi \to \psi) \not\vdash_C (\mathtt{B}\varphi \to \mathtt{B}\psi) \qquad \text{(Fake K)}$$

$$\mathtt{B}(\varphi \to \psi), \mathtt{B}\varphi \vdash_C \mathtt{B}\psi \qquad \text{(Real K)}$$

The answer is, however, that putting $\mathtt{B}\varphi \to \mathtt{B}\psi$ on the right-hand side of $\vdash_C$ doesn't adequately capture the intuition that *if φ is informally provable, then so is ψ*. For $\mathtt{B}\varphi \to \mathtt{B}\psi$ actually contains more information than that. *If φ is informally provable, then so is ψ* tells us only what happens when φ (and so, $\mathtt{B}\varphi$) is informally provable, while the provability of $\mathtt{B}\varphi \to \mathtt{B}\psi$ puts further constraints on what happens if φ is not informally provable (for instance, that if it is undecidable, ψ cannot be refutable). It's (Real K) and not (Fake K) that properly captures the underlying intuition.

Given our discussion of implication and the failure of the deduction theorem, one key observation is that we have:

$$\mathtt{B}\varphi \vdash_C \varphi \qquad \text{(Cref)}$$

which is a certain version of the reflection schema.

6.7 CABAT and provability paradoxes

Now, let's see one important difference between using $\vdash_C$ and its provability operator on the one hand, and using Peano Arithmetic and its standard provability predicate (or the modal logic of provability GL and its provability operator, for that matter) on the other.

Quite crucially, we may want to see which principles that hold for standard formal provability predicates hold for the operator $\mathtt{B}$ as well. In the arithmetical setting we get Löb's theorem as a side-effect of Diagonal Lemma. It is not something that we would like to postulate as an interesting and independently motivated principle. Rather, it is an unwanted surprising consequence. It is also one of the reasons why we cannot consistently put all the instances of the reflection schema together with HB conditions in the classical setting.

In CABAT we have certain versions of HB conditions:

$$\varphi \vdash_C \mathtt{B}\varphi \qquad \text{(HB1}')$$

$$\mathtt{B}(\varphi \to \psi), \mathtt{B}\varphi \vdash_C \mathtt{B}\psi \qquad \text{(HB2}')$$

$$\mathtt{B}\varphi \vdash_C \mathtt{BB}\varphi \qquad \text{(HB3}')$$

The first condition in CABAT is a bit stronger than (HB1), since it is not restricted only to theorems. The condition starts to be intuitive as soon as you recall the interpretation of $\varphi \vdash_C \psi$, which says that if φ is informally provable then so is ψ. Some may be worried that the above formulation of HB1, in some sense, allows to go from premises which are true (and may not be theorems) to premises which are theorems. But as we explained, according to our reading formulas on the left hand side of $\vdash_C$ are not true but informally provable. So the principle allows only to go from informally provable premises to informally provable premises having informal provability expressed in the object language.

CABAT doesn't validate reflection as implication in its whole generality. Given that implication has a rather strong meaning, as already discussed, we don't consider it a flaw. Our intended formulation of reflection is (Cref), which we do have. Nevertheless, we have the following theorem.

Theorem 99 (CABAT reflection as implication)**.** *Suppose either* $\vdash_C \varphi$ *or* $\vdash_C \neg\varphi$*. It follows that* $\vdash_C \mathtt{B}\varphi \to \varphi$*.*

Proof. Take any evaluation e, it is clear that since $e(\varphi) = 1$ or $e(\varphi) = 0$, we have $e(\neg\mathtt{B}\varphi \vee \varphi) = 1$, which shows that $e(\mathtt{B}\varphi \to \varphi) = 1$. □

Given that (Nec) is admissible in CABAT, having full reflection as implication would lead to inconsistency as it did for the formal provability predicate.

Observe an interesting corollary of Theorem 99: CABAT is not comparable to the logic of formal provability GL (see: (Boolos, 1993)). On the one hand, CABAT proves reflection for countertautologies (GL doesn't) and on the other hand $\mathtt{GL} \vdash \mathtt{B}(\varphi \to \psi) \to (\mathtt{B}\varphi \to \mathtt{B}\psi)$ which is not a valid schema in CABAT.

An interesting question is whether the inferential apparatus of CABAT is sufficient to prove Löb's theorem. The key observation is that in the standard proof, once the premises, including the one produced by an application of the diagonal lemma, are listed, the theorem follows by classical propositional logic. So the issue can be judged at the propositional level, once all the relevant assumptions are in.

The natural way to go about the translation is this. We translate both $\mathtt{Pr}$ and $\vdash$ as $\mathtt{B}$. It is a standard practice to translate them using a single symbol (see Boolos, 1993). Slightly more challenging is the question how to translate implications from the language of **PA**. The straightforward approach is to translate them as material implications in $\mathrm{L}_{\mathtt{B}}$. But we think it will not do justice to the original theorem. The deduction theorem does not hold for CABAT. Implications are stronger claims than consequence claims and are much harder to prove. Thus, whenever possible, we will translate $\varphi \to \psi$ in the conclusions as $\varphi \vdash_C \psi$. We leave implications in the premises, especially within the scope of $\mathtt{B}$. But this is not a cheap way for us to avoid an undesired consequence: by leaving material implications in the premises we make them as strong as we can. As for sentences produced by the application of Diagonal Lemma we will build them to the assumptions.

Fact 100 (Löb's theorem failure)**.**

$$\mathtt{B}(\mathtt{B}\varphi \to \varphi), \mathtt{B}(\lambda \to (\mathtt{B}\lambda \to \varphi)), \mathtt{B}((\mathtt{B}\lambda \to \varphi) \to \lambda) \not\vdash_C \mathtt{B}\varphi.$$

Proof. Just take an assignment $v(\varphi) = v(\lambda) = n$ and extend it to an evaluation where for each implication if it is possible to choose n, it should be chosen. It is easily verifiable that all the premises have value 1, and yet the conclusion has value n, while all the constraints on valuations remain satisfied.[28] □

Why doesn't the standard argument work? Suppose e_v gives value 1 to all the premises. Since $e_v(\mathtt{B}(\lambda \to (\mathtt{B}\lambda \to \varphi))) = 1$, it follows that $e_v(\mathtt{B}\lambda \to (\mathtt{BB}\lambda \to \mathtt{B}\varphi)) = 1$. Now, in the standard proof we use the fact that $e_v(\mathtt{B}\varphi \to \mathtt{BB}\varphi) = 1$, but we cannot do that here, since in general the previous formula is not a CABAT-tautology.

[28] Note that here we are translating the assumptions of Löb's theorem not adding reflection. As soon as we add instance of reflection for $\neg\lambda$ the unwanted consequence holds.

In other words, it is not the case that only those instances of the reflection schema are provable for which φ is already a theorem. The lack of Löb's theorem is rather promising, since it isn't too intuitive for informal provability. In fact, we've seen that reflection as implication is provable also for refutable claims, and reflection as consequence is in for all formulas.

We will take a quick look at two other theorems related to provability and the reflection schema. Montague theorem suggested that it is impossible to add all instances of the schema and at the same time have all the Hilbert-Bernays conditions. The moral from Montague (and dual Montague) theorems is that the price for all Hilbert-Bernays conditions together with all instances of the reflection schema on the one hand, or all instances of provabilitation one the other (assuming co-necessitation) is too high in the classical setting. But what happens when we switch to CABAT?

First, note that every occurrence of $\mathtt{B}$ on the left-hand side of $\vdash_C$ can be omitted. Also, for the clarity we split equivalences into implications. The relevant inferences, when evaluated from the perspective of CABAT, result in the following.

Fact 101 (Truth-teller)**.** $\lambda \to \mathtt{B}\lambda, \mathtt{B}\lambda \to \lambda \not\vdash_C \lambda$.

Proof. Take any evaluation e which ascribes to all the premises value 1. It is possible that e gives value n to both $\lambda, \mathtt{B}\lambda$. Under this evaluation the conclusion doesn't have value 1.

□

Fact 102 (Curry)**.** $\lambda \to (\lambda \to \varphi), (\lambda \to \varphi) \to \lambda \vdash_C \varphi$.

Proof. The reasoning is classically valid, so it is also valid in CABAT. □

Fact 103 (Montague with IPL)**.**

$$\lambda \to \mathtt{B}\neg\lambda, \mathtt{B}\neg\lambda \to \lambda, \mathtt{B}\lambda \to \lambda, \mathtt{B}\neg\lambda \to \neg\lambda \vdash_C \lambda \wedge \neg\lambda.$$

Proof. Take an evaluation e such that all the premises have value 1. Clearly, since $e(\lambda \to \mathtt{B}\neg\lambda) = 1 = e(\mathtt{B}\neg\lambda \to \neg\lambda)$, it follows that $e(\lambda) = 0$. On the other hand $e(\mathtt{B}\neg\lambda \to \lambda) = 1$ so $e(\neg\mathtt{B}\neg\lambda) = 1$, which means $e(\mathtt{B}\neg\lambda) = 0$, which implies $e(\lambda) = 1$. □

The moral is quite straightforward. We blocked Löb's theorem but not at a very high price: other counterparts of formal results that hold in the classical context still work. Just for the sake of completeness we will take a glance at the behavior of CABAT-counterparts of other theorems.

Fact 104 (Dual Montague with IPL)**.** $\lambda \to \mathtt{B}\neg\lambda, \mathtt{B}\neg\lambda \to \lambda, \lambda \to \mathtt{B}\lambda, \neg\lambda \to \mathtt{B}\neg\lambda \vdash_C \lambda \wedge \lambda$.

Proof. Take any evaluation e which gives value 1 to all the premises. Clearly, since $e(\neg\lambda \to \mathtt{B}\neg\lambda) = 1 = e(\lambda \to B\neg\lambda)$ it follows that $e(\mathtt{B}\neg\lambda) = 1$ so $e(\lambda) = 0$. On the other hand $e(\mathtt{B}\neg\lambda \to \lambda) = 1$ so $e(\mathtt{B}\neg\lambda) = 1$, thus $e(\lambda) \neq 0$, which is a contradiction. □

What is also interesting is the fact that the dual paradox in which we add provabilitation instead of reflection still works.

Fact 105 (Montague with IPG)**.** *The following consequence still holds:*

$$\lambda \to \neg \mathtt{B}\lambda, \neg \mathtt{B}\lambda \to \lambda, \mathtt{B}\lambda \to \lambda, \mathtt{B}\neg\lambda \to \neg\lambda \nvdash_C \lambda \wedge \neg\lambda.$$

Proof. Take any evaluation e which gives value 1 to all the premises. Clearly, since $e(\lambda \to \neg \mathtt{B}\lambda) = 1 = e(\neg \mathtt{B}\lambda \to \neg\lambda)$, it follows $e(\lambda \to \neg\lambda) = 1$, so $e(\neg\lambda) = 1$. On the other hand $e(\neg \mathtt{B}\lambda \to \lambda) = 1$, which implies that $e(\mathtt{B}\lambda) = 1$ so $e(\lambda) = 1$ a contradiction. □

Fact 106 (Dual Montague with IPG)**.** *The following consequence still holds:*

$$\lambda \to \neg \mathtt{B}\lambda, \neg \mathtt{B}\lambda \to \lambda, \lambda \to \mathtt{B}\lambda, \neg\lambda \to \mathtt{B}\neg\lambda \nvdash_C \lambda \wedge \neg\lambda.$$

Proof. Take any evaluation e which gives value 1 to all the premises. Clearly, since $e(\neg\lambda \to \mathtt{B}\neg\lambda) = 1 = e(\neg \mathtt{B}\lambda \to \lambda)$, it follows that $e(\lambda) = 1$. On the other hand, $e(\lambda \to \neg \mathtt{B}\lambda) = 1$, so $e(\neg \mathtt{B}\lambda) = 1$ which means $e(\lambda) \neq 1$. □

Given these results, we submit, CABAT, despite the lack of reflection formulated by means of implication, incorporates reflection nevertheless, and is an independently motivated and interesting candidate to be further developed into a first order version. It incorporates very basic intuitions about informal provability as used in mathematical practice and preserves quite a few intuitive inferential patterns of informal provability at a not too high cost.

Chapter 7

Future work and conclusions

First, let's say a few words about possible dimensions in which the framework can be further extended. We start with a first order version of BAT.

7.1 First order BAT

In order to construct a first order version of BAT, we need to start with a couple of definitional and notational conventions. First, let $\mathcal{L}$ be a first order language understood as a set of formulas built in the standard way. We use $Var = x_1, x_2, \ldots$ to denote the set of variables, Con for constants, $Term$ for the set of terms. Sometimes we will be interested in the language $\mathcal{L}^+$ defined as $\mathcal{L}$ extended with constants for all elements of the quantification domain. Usually, to define a three-valued first order logic the notion of a three-valued structure is used. This notion is pretty standard (Halbach, 2011). In our case, we use a slight variation of this notion, since the resulting logic is not deterministic. It's quite clear how to rephrase the results in the paper in terms of the standard three-valued setting.

A *three-valued structure* is a tuple $\langle \mathtt{M}, i \rangle$, such that:

- $\mathbb{M} \neq \emptyset$ is the domain of quantification (sometimes called the universe of the structure).
- i is an interpretation:
  - To every n-ary predicate P, i ascribes a triple $\langle E(P), A(P), F(P) \rangle$ such that:

$$E(P), A(P), F(P) \subseteq \mathtt{M}^n$$
$$E(P) \cap A(P) = E(P) \cap F(P) = A(P) \cap F(P) = \emptyset$$
$$E(P) \cup F(P) \cup A(P) = \mathtt{M}^n$$

    $E(P)$ is called the *extension* of a predicate P, $A(P)$ stands for the *anti-extension* of P and $F(P)$ is called the *fringe* of P. Intuitively, $E(P)$ corresponds to the things that are P, $A(P)$ to the things that are not P and the fringe correspond to those elements of the domain to which a predicate is not applicable. In the classical context we always assume that the fringe is empty.

- $i(c) \in \mathtt{M}$ for every constant c.
- For any n-ary function symbol $\circ$, $i(\circ) : \mathtt{M}^n \mapsto \mathtt{M}$.
- Identity is classical: $i(=)$ is $\langle E(=), A(=), F(=)\rangle$ such that $E(=)$ is $\{\langle x, x\rangle \mid x \in \mathtt{M}\}$, $A(=)$ is $\mathtt{M}^2 \setminus E(=)$ and $F(=)$ is empty.

Now, a *three-valued model* $\mathcal{M}$ is a triple $\langle \mathtt{M}, i, v\rangle$, where $\langle \mathtt{M}, i\rangle$ is a three-valued structure and $v : Var \mapsto \mathtt{M}$ is a valuation function. Relative to a valuation we can define the interpretation of terms:

- $t^{\mathcal{M}}(\tau) = i(\tau)$ if τ is a constant,
- $t^{\mathcal{M}}(x) = v(x)$ if $x \in Var$,
- $t^{\mathcal{M}}(\circ(\tau_1, \ldots \tau_n)) = (i(\circ))(t^{\mathcal{M}}(\tau_1), \ldots, t^{\mathcal{M}}(\tau_n))$.

For a moment let's focus on atomic formulas. In the classical context each atomic formula $P(a)$ is either true if $i(a) \in E(P)$ or it's false, if $i(a) \in A(P)$. The interpretation uniquely determines the values of all formulas defining at the same time the satisfaction relation. In our case, since we have three values, it's impossible to use the classical satisfaction relation. Instead, we use a triple $\langle$  ,  ,  $\rangle$ of relations between a given structure $\mathcal{M}$ and the set of well-formed formulas partially determined by:

- $\mathcal{M}$  $P(\tau_1, \ldots \tau_n)$ iff $\langle i(\tau_1), \ldots, i(\tau_n)\rangle \in E(P)$
- $\mathcal{M}$  $P(\tau_1, \ldots, \tau_n)$ iff $\langle i(\tau_1), \ldots, i(\tau_n)\rangle \in A(P)$
- $\mathcal{M}$ $P(\tau_1, \ldots, \tau_n)$ iff $\langle i(\tau_1), \ldots, i(\tau_n)\rangle \in F(P)$

These *satisfaction triples* are used to define the notion of *evaluation* which is responsible for generating the satisfaction clauses for the full language. *An evaluation* is a total function $e_{\mathcal{M}} : \mathcal{L} \mapsto \{0, n, 1\}$ such that for atomic formulas φ we have:

- $e_{\mathcal{M}}(\varphi) = 1$ iff $\mathcal{M}$ φ,
- $e_{\mathcal{M}}(\varphi) = n$ iff $\mathcal{M}$ φ,
- $e_{\mathcal{M}}(\varphi) = 0$ iff $\mathcal{M}$ φ.

Again, in classical two or three-valued settings evaluations are uniquely determined by a model and truth-tables for connectives and satisfaction clauses of quantifiers. Since, our aim is to generalize BAT to the first order, we need to restrict our attention to those evaluations which satisfy conditions put on connectives by BAT. In order to cope with quantifiers, we treat them as "infinite" conjunction and infinite disjunction.

Definition 107 (BAT evaluation)**.** Let $\mathcal{M}$ be a three-valued model. We say that an evaluation $e_{\mathcal{M}}$ is a *BAT evaluation* iff for all formulas φ, ψ:

- $e_{\mathcal{M}}(\neg\varphi) = 1$ iff $e_{\mathcal{M}}(\varphi) = 0$,

- $e_{\mathcal{M}}(\neg\varphi) = n$ iff $e_{\mathcal{M}}(\varphi) = n$,
- $e_{\mathcal{M}}(\neg\varphi) = 0$ iff $e_{\mathcal{M}}(\varphi) = 1$,
- If $e_{\mathcal{M}}(\varphi) = 1$ or $e_{\mathcal{M}}(\psi) = 1$, then $e_{\mathcal{M}}(\varphi \vee \psi) = 1$,
- $e_{\mathcal{M}}(\varphi \vee \psi) = 0$ iff $e_{\mathcal{M}}(\varphi) = 0$ and $e_{\mathcal{M}}(\psi) = 0$,
- If $e_{\mathcal{M}}(\varphi) = 0$ and $e_{\mathcal{M}}(\psi) = n$, then $e_{\mathcal{M}}(\varphi \vee \psi) = n$,
- If $e_{\mathcal{M}}(\varphi) = n$ and $e_{\mathcal{M}}(\psi) = 0$, then $e_{\mathcal{M}}(\varphi \vee \psi) = n$,
- If $e_{\mathcal{M}}(\varphi) = n$ and $e_{\mathcal{M}}(\psi) = n$, then $e_{\mathcal{M}}(\varphi \vee \psi) = n$ or $e_{\mathcal{M}}(\varphi \vee \psi) = 1$,
- If $e_{\mathcal{M}}(\varphi) = 0$ or $e_{\mathcal{M}}(\psi) = 0$, then $e_{\mathcal{M}}(\varphi \wedge \psi) = 0$,
- $e_{\mathcal{M}}(\varphi \wedge \psi) = 1$ iff $e_{\mathcal{M}}(\varphi) = 1$ and $e_{\mathcal{M}}(\psi) = 1$,
- If $e_{\mathcal{M}}(\varphi) = 1$ and $e_{\mathcal{M}}(\psi) = n$, then $e_{\mathcal{M}}(\varphi \wedge \psi) = n$,
- If $e_{\mathcal{M}}(\varphi) = n$ and $e_{\mathcal{M}}(\psi) = 1$, then $e_{\mathcal{M}}(\varphi \wedge \psi) = n$,
- If $e_{\mathcal{M}}(\varphi) = n$ and $e_{\mathcal{M}}(\psi) = n$, then $e_{\mathcal{M}}(\varphi \wedge \psi) = n$ or $e_{\mathcal{M}}(\varphi \wedge \psi) = 0$,
- $e_{\mathcal{M}}(\varphi \to \psi) = 0$ iff $e_{\mathcal{M}}(\psi) = 1$, then $e_{\mathcal{M}}(\varphi \wedge \psi) = 0$,
- If $e_{\mathcal{M}}(\varphi) = 0$ then $e_{\mathcal{M}}(\varphi \to \psi) = 1$,
- If $e_{\mathcal{M}}(\varphi) = n$ and $e_{\mathcal{M}}(\psi) = n$, then $e_{\mathcal{M}}(\varphi \to \psi) = n$ or $e_{\mathcal{M}}(\varphi \to \psi) = 1$,
- If $e_{\mathcal{M}}(\varphi) = n$ and $e_{\mathcal{M}}(\psi) = 0$, then $e_{\mathcal{M}}(\varphi \to \psi) = n$,
- If $e_{\mathcal{M}}(\varphi) = 1$ and $e_{\mathcal{M}}(\psi) = n$, then $e_{\mathcal{M}}(\varphi \to \psi) = n$,
- If $e_{\mathcal{M}}(\varphi) = 1$ and $e_{\mathcal{M}}(\psi) = 1$, then $e_{\mathcal{M}}(\varphi \to \psi) = 1$,
- If $e_{\mathcal{M}}(\varphi) = n$ and $e_{\mathcal{M}}(\psi) = 1$, then $e_{\mathcal{M}}(\varphi \to \psi) = 1$,
- The clauses for equivalence are exactly as those of $(\varphi \to \psi) \wedge (\psi \to \varphi)$,
- $e_{\mathcal{M}}(\forall x\, \varphi(x)) = 1$ iff for any constant $c \in \mathcal{L}^+$, we have $e_{\mathcal{M}}(\varphi(c)) = 1$,
- $e_{\mathcal{M}}(\forall x\, \varphi(x)) = 0$ iff there is a constant $c \in \mathcal{L}^+$ such that $e_{\mathcal{M}}(\varphi(c)) = 0$,
- $e_{\mathcal{M}}(\forall x\, \varphi(x)) = 0$ or $e_{\mathcal{M}}(\forall x\, \varphi(x)) = n$ otherwise,
- $e_{\mathcal{M}}(\exists x\, \varphi(x)) = 0$ iff for all constants $c \in \mathcal{L}^+$ such that $e_{\mathcal{M}}(\varphi(c)) = 0$,
- $e_{\mathcal{M}}(\exists x\, \varphi(x)) = 1$ iff there is a constant $c \in \mathcal{L}^+$ such that $e_{\mathcal{M}}(\varphi(c)) = 1$,
- $e_{\mathcal{M}}(\exists x\, \varphi(x)) = n$ or $e_{\mathcal{M}}(\exists x\, \varphi(x)) = 1$ otherwise.

The above clauses limit the set of all possible evaluations to BAT-evaluations. The intuition here is quite simple: we try to mimic the clauses from the propositional level. Quantifiers are treated as a generalizations of BAT conjunction and disjunction. In a sense quantifiers are "infinitary" versions of them. Each such a BAT-evaluation determines a *BAT triple* $\langle \blacktriangledown_B, \blacktriangle_B, \blacklozenge_B \rangle$ which is an extension of a satisfaction triple to the full language. Based on that, we can define the following notions:

Definition 108 (BAT-triple based on $\mathcal{M}$)**.** Let $\mathcal{M}$ be a model. We say that $\langle \blacktriangledown, \blacktriangle, \blacklozenge \rangle$ is *based on* $\mathcal{M}$ iff $\langle \blacktriangledown, \blacktriangle, \blacklozenge \rangle$ is a BAT-triple defined over the same model. The set of all BAT-triples based on $\mathcal{M}$ is denoted as $Str_{\mathcal{M}}$. We say that a model $\mathcal{M}$ decides formula φ iff all BAT-evaluations agree on φ, so all BAT-evaluations ascribe the same value to φ.

Every model decides all Boolean combinations of atoms which are already decided by it, but it is quite easy to see that in general $\mathcal{M}$ does not decide the logical values of all complex formulas. For instance, for disjunctions of atoms which are undecided by the model there are many underlying BAT-triples.

We managed to lift the framework to fully-fledged first order. Unfortunately, the whole framework is quite complex and some philosophically motivated technical issues remain unsettled. For instance, given a three-valued structure it is not clear what is the structure of BAT-evaluations based on it. Our hypothesis is that they form a lattice with maximal and minimal elements, where the order between BAT-evaluations is introduced by looking at the formulas whose values are 1. The next issue is to find a procedure for selecting the BAT-evaluations that are "admissible" to parallel the propositional move from BAT to CABAT. Note that if we consider the set of all evaluations based on a given structure, then we have a similar problem as in the propositional case: disjunction is not symmetric. Thus, there is a need for an additional closure or a filtration condition which would allow one to eliminate the unwanted BAT evaluations. The third issue is to actually use the framework over arithmetic and see what happens.

7.2 Conclusions

The non-deterministic approach to informal provability is far from complete. It is rather clear, from the thesis, that the paradigm is quite promising and it is a source of interesting technical results. We showed that CABAT is a viable candidate as the logic of informal provability. With CABAT in the background we have more reflection: it is no longer limited only to theorems.

Not only the study of non-deterministic semantics is interesting but also the relation between well-known many-valued logics and the non-deterministic semantics is far from clear. Non-deterministic semantics is difficult to work with. One of the problems with it, seems to stem from its generality. This makes the investigations about alternative semantics for the logic of informal provability attractive.

Philosophically, I showed that the non-deterministic interpretation is far from insane. It is well-motivated and influenced by a reading of informal provability quite close to mathematical practice. To be fair, the cost of a non-deterministic

solution is the technical complexity and the lack of natural semantics for CABAT. As I showed, CABAT cannot be pinned down by most of known semantics that are usually used to characterize propositional modal logics. Yet, it seems that the principles of informal provability incorporated in CABAT are sufficient to obtain a strong theory which somehow is not susceptible to most of the well-known paradoxes. Note that the relation between formal and formal provability is quite complex. On the one hand, it seems that inferential patterns governing these two notions are different. But is it enough to infer that the extensions of these concepts are different? I think, there is still room for an interesting philosophical debate. On the technical side, it may be interesting to develop a system containing two provability predicates both formal and informal to see now how they interact in the first order arithmetical setting.

I also discussed Löb's theorem and its lack of applicability to the informal concept of a proof. It is one of the reasons blocking the straightforward strategy for adding the reflection schema together with others intuitive principles of informal provability. Philosophically, it does not constitute a convincing principle, since it's not compatible with the reflection schema. In CABAT Löb's theorem is not provable, and so the amount of the reflection that CABAT admits is greater than for the formal provability.

Finally, I was able to study the most common types of the arguments to the effect that informal mathematics, taken as a whole, cannot be consistent. The construction of CABAT, and philosophical discussion of thereof shows that it is possible to have an interesting formalization of informal provability on top of which one can argue against the inconsistency of informal mathematics.

Bibliography

Aczel, P. (1988). *Non-well-founded sets*. Center for the Study of Language and Information, Stanford, CA.

Antonutti Marfori, M. (2010). Informal proofs and mathematical rigour. *Studia Logica*, 96:261–272.

Arai, T. (1998). Some results on cut-elimination, provable well-orderings, induction and reflection. *Annals of Pure and Applied Logic*, 95(1-3):93–184.

Artemov, S. (1994). Logic of proofs. *Annals of Pure and Applied Logic*, 67(1-3):29–59.

Artemov, S. (1998). Logic of proofs: a unified semantics for modality and λ-terms. Technical report, Cornell University, CFIS 98–06.

Avron and Zamanski (2011). Non-deterministic semantics for logical systems. In Gabbay and Guenthner, editors, *Handbook of Philosophical Logic*, volume 16. Springer Netherlands.

Avron, A. and Lev, I. (2001). Canonical propositional gentzen-type systems. In *Automated Reasoning: First International Joint Conference, IJCAR 2001 Siena, Italy, June 18-23, 2001 Proceedings (Lecture Notes in Computer Science)*, pages 529–544. Springer-Verlag.

Avron, A. and Lev, I. (2005). Non-deterministic multiple-valued structures. *Journal of Logic and Computation*, 15(3):241–261.

Baaz, M., Fermüller, C. G., Salzer, G., and Zach, R. (1996). Multlog 1.0: Towards an expert system for many-valued logics. In *International Conference on Automated Deduction*, pages 226–230. Springer.

Batens, D. (1998). A dynamic semantics for inconsistency-adaptive logics. *Bulletin of the Section of Logic*, 27(15-8):5l.

Batens, D. (1999). Inconsistency-adaptive logics. In *Logic at Work. Essays Dedicated to the Memory of Helena Rasiowa*, pages 445–472. Springer.

Batens, D. (2000). A survey of inconsistency-adaptive logics. In *Frontiers of paraconsistent logic*, pages 49–73. Research Studies Press.

Batens, D. and De Clercq, K. (2004). A rich paraconsistent extension of full positive logic. *Logique et Analyse*, 185-188.

Beall, J. (1999). Prom full blooded platonism to really full blooded platonism. *Philosophia Mathematica*, 7(3):322–325.

Beklemishev, L. D. (1997). Induction rules, reflection principles, and provably recursive functions. *Annals of Pure and Applied Logic*, 85(3):193–242.

Beklemishev, L. D. (2003). Proof-theoretic analysis by iterated reflection. *Archive for Mathematical Logic*, 42(6):515–552.

Belnap, N. D. (1962). Tonk, plonk and plink. *Analysis*, 22(6):130–134.

Bergmann, M. (2008). *An introduction to many-valued and fuzzy logic: semantics, algebras, and derivation systems*. Cambridge University Press, Cambridge New York.

Beth, E. (1955). Semantic entailment and formal derivability. *Mededelingen van de Koninklijke Nederlandse Akademie van Wetenschappen*, 18(13):309–342.

Blackburn, P. (2001). *Modal logic*. Cambridge University Press, Cambridge England New York.

Bochvar, D. A. (1939). On a three valued calculus and its application to the analysis of contradictories. *Matematicheskii Sbornik*, 4(2):287–308.

Boolos, G. (1993). *The Logic of Provability*. Cambridge University Press.

Carnielli, W. and Matulovic, M. (2015). The method of polynomial ring calculus and its potentialities. *Theoretical Computer Science*, 606:42–56.

Carnielli, W. A. (1987). Systematization of finite many-valued logics through the method of tableaux. *The Journal of Symbolic Logic*, 52(2):473–493.

Chellas, B. (1980). *Modal logic : an introduction*. Cambridge University Press, Cambridge England New York.

Church, A. (1996). *Introduction to mathematical logic*. Princeton University Press, Princeton, N.J.

Coniglio, M. E., Fariñas del Cerro, L., and Peron, N. M. (2015). Finite non-deterministic semantics for some modal systems. *Journal of Applied Non-Classical Logics*, 25(1):20–45.

Coniglio, M. E. and Peron, N. M. (2014). Dugundji's theorem revisited. *Logica Universalis*, 8(3-4):407–422.

Dugundji, J. (1940). Note on a property of matrices for lewis and langford's calculi of propositions. *Journal of Symbolic Logic*, 5(4):150–151.

Enderton, H. (1977). *Elements of set theory*. Academic Press, New York.

Feferman, S., Dawson Jr, J. W., Kleene, S. C., Moore, G. H., Solovay, R. M., and van Heijenoort, J., editors (1986). *Kurt Gödel: collected works. Vol. 1: Publications 1929-1936*. Oxford University Press.

Fitting, M. (2003). A semantics for the logic of proofs. Technical report, CUNY Ph. D. Program in Computer Science, TR - 2003012.

Flagg, R. C. and Friedman, H. (1986). Epistemic and intuitionistic formal systems. *Annals of Pure and Applied Logic*, 32(1):53–60.

Franzén, T. (2004). *Inexhaustibility: a non-exhaustive treatment*. Association for Symbolic Logic A K Peters, Urbana, Ill. Wellesley, Mass.

Gödel, K. (1953). Is mathematics syntax of language? In *K. Gödel Collected Works*, pages 334–355. Oxford University Press: Oxford.

Gödel, K. (1986). An interpretation of the intuitionistic propositional calculus. *Collected Works*, 1:301–303.

Goodman, N. D. (1984). Epistemic arithmetic is a conservative extension of intuitionistic arithmetic. *Journal of Symbolic Logic*, 49(1):192–203.

Hájek, P. and Pudlak, P. (1993). Metamathematics of first-order arithmetic. *Berlin: Springer*, 2:295–297.

Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambridge University Press.

Halbach, V. and Horsten, L. (2000). Two proof-theoretic remarks on EA + ECT.

Mathematical Logic Quarterly, 46(4):461–466.

Halbach, V. and Visser, A. (2014). Self-reference in arithmetic i. *Review of Symbolic Logic*, 7(4):671–691.

Henkin, L. (1952). A problem concerning provability. *Journal of Symbolic Logic*, 17(2):160.

Heylen, J. (2013). Modal-epistemic arithmetic and the problem of quantifying in. *Synthese*, 190(1):89–111.

Hilbert, D. and Bernays, P. (1939). *Grundlagen der Mathematik II*. Springer.

Horsten, L. (1994). Modal-epistemic variants of Shapiro's system of epistemic arithmetic. *Notre Dame Journal of Formal Logic*, 35(2):284–291.

Horsten, L. (1996). Reflecting in epistemic arithemtic. *The Journal of Symbolic Logic*, 61:788–801.

Horsten, L. (1997). Provability in principle and controversial constructivistic principles. *Journal of Philosophical Logic*, 26(6):635–660.

Horsten, L. (1998). In defence of epistemic arithmetic. *Synthese*, 116:1–25.

Horsten, L. (2002). An axiomatic investigation of provability as a primitive predicate. In *Principles of Truth*, pages 203–220. Hansel-Hohenhausen.

Horsten, L. (2011). *The Tarskian Turn. Deflationism and Axiomatic Truth*. Mit Press.

Koellner, P. (2016). Gödel's disjunction. In Horsten, L. and Welch, P., editors, *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge*. Oxford University Press UK.

Kripke, S. A. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72(19):690–716.

Leitgeb, H. (2009). On formal and informal provability. In *New Waves in Philosophy of Mathematics*, pages 263–299. New York: Palgrave Macmillan.

Löb, M. H. (1955). Solution of a problem of Leon Henkin. *The Journal of Symbolic Logic*, 20(02):115–118.

Łukasiewicz, J. (1970). *Selected Works*. Amsterdam: North-Holland Pub. Co.

Łukasiewicz, J. (1988). O logice trójwartościowej. *Studia Filozoficzne*, 270(5).

Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta philosophica Fennica*, (16):153–167.

Myhill, J. (1960). Some remarks on the notion of proof. *Journal of Philosophy*, 57(14):461–471.

Omori and Skurt (2016). More modal semantics without possible worlds. *IfCoLog Journal of Logics and their Applications*, (3):815–845.

Pawlowski, P. and Urbaniak, R. (2017). Many-valued logic of informal provability: a non-deterministic strategy [accepted and forthcoming]. *The Review of Symbolic Logic*.

Priest, G. (2001). *Introduction to Non-Classical Logic*. Cambridge University Press.

Priest, G. (2006). *In Contradiction*. Oxford University Press UK.

Quine, W. V. (1974). *The Roots of Reference*. Lasalle, Ill., Open Court.

Rav, Y. (1999). Why do we prove theorems? *Philosophia Mathematica*, 7(1):5–41.

Rav, Y. (2007). A critique of a formalist-mechanist version of the justification of arguments in mathematicians' proof practices. *Philosophia Mathematica*, 15(3):291–320.

Reinhardt, W. N. (1986). Epistemic theories and the interpretation of Gödel's incompleteness theorems. *Journal of Philosophical Logic*, 15(4):427–74.

Rin, B. G. and Walsh, S. (2016). Realizability semantics for quantified modal logic: Generalizing Flagg's 1985 construction. *The Review of Symbolic Logic*, 9(4):752–809.

Segerberg, K. K. (1971). *An essay in classical modal logic*. The Philosophical Society in Uppsala.

Shapiro, S. (1985). Epistemic and intuitionistic arithemtic. In *Intensional mathematics*. North Holland.

Sjögren, J. (2010). A note on the relation between formal and informal proof. *Acta Analytica*, 25(4):447–458Mo.

Smith, P. (2007). *An Introduction to Gödel's Theorems*. Cambridge University Press.

Solovay, R. (1976). Provability interpretations of modal logic. *Israel Journal of Mathematics*, 25.

Stern, J. (2015). *Toward Predicate Approaches to Modality*. Trends in Logic. Springer.

Tanswell, F. (2015). A problem with the dependence of informal proofs on formal proofs. *Philosophia Mathematica*, 23(3):295–310.

Tanswell, F. (2016). Saving proof from paradox: Gödel's paradox and the inconsistency of informal mathematics. In Andreas, H. and Verdee, P., editors, *Logical Studies of Paraconsistent Reasoning in Science and Mathematics*. Springer.

Urbaniak, R. and Pawlowski, P. (2018). Logics of (formal and informal) provability [accepted and forthcoming]. In Hansson, S. O., Hendrick, V., and Michelsen, K. E., editors, *Introduction to formal philosophy*. Springer.

Summary

Mathematicians prove theorems in a semi-formal setting, providing what we'll call *informal proofs*. There are various philosophical reasons not to reduce informal provability to formal provability within some appropriate axiomatic theory, but the main worry is that we seem committed to all instances of the so-called reflection schema: $\mathtt{B}(\varphi) \to \varphi$ (where $\mathtt{B}$ stands for the informal provability predicate). Yet, adding all its instances to any theory for which Löb's theorem ($\mathtt{B}(\mathtt{B}\varphi \to \varphi) \to \mathtt{B}(\varphi)$) and some other very intuitive conditions for $\mathtt{B}$ hold leads to inconsistency.

We propose a new way out by treating informal provability as a partial notion. This means that some mathematical sentences are informally provable, some are informally refutable and some are neither. In order to model this formally, we develop the three-valued non-deterministic logics BAT and CABAT. The reason for the lack of truth-functionality stems from the observation that disjunctions and conjunctions of sentences which are neither informally provable nor refutable may have a different status depending on the relations between these sentences.

Most of the dissertation is devoted to an extensive study of these logics. In particular, we search for a less complex semantics, showing that many natural candidates (finite deterministic semantics, Kripke semantics, non-normal Kripke semantics and neighbourhood semantics) are not up for the job. We also construct a complete proof system and generalize it to any non-deterministic consequence relation. We also study some of the paradoxes of informal provability, investigating what happens with them if we change the underlying logic from classical logic to CABAT.

Samenvatting

Wiskundigen bewijzen stellingen in een semi-formeel kader en komen op die manier tot wat "informele bewijzen" worden genoemd. Er zijn verschillende filosofische redenen om informele bewijsbaarheid niet te reduceren tot formele bewijsbaarheid binnen een axiomatische theorie. De voornaamste is dat het niet te verantwoorden lijkt om niet alle instanties van het zogenaamde reflectieschema ($\mathtt{B}\varphi \to \varphi$) toe te voegen (waarbij $\mathtt{B}$ staat voor het informele bewijsbaarheidspredikaat). Er kan echter worden aangetoond dat het toevoegen van alle instanties van het reflectieschema aan een theorie, waarin zowel de stelling van Löb ($\mathtt{B}(\mathtt{B}\varphi \to \varphi) \to \mathtt{B}(\varphi)$) geldt als een aantal andere intuïtieve eigenschappen voor $\mathtt{B}$, onvermijdelijk tot inconsistenties leidt.

In dit proefschrift stellen we een nieuwe manier voor om hiermee om te gaan, namelijk door informele bewijsbaarheid als een partiële notie te beschouwen. Dit komt erop neer dat we sommige beweringen als informeel bewijsbaar beschouwen, sommige als informeel weerlegbaar en sommige als geen van beide. Om dit formeel vorm te geven, hebben we de driewaardige niet-deterministische logica's BAT en CABAT ontworpen. Dat deze logica's niet waarheidsfunctioneel zijn, hangt samen met de vaststelling dat disjuncties en conjuncties van zinnen die noch informeel bewijsbaar noch informeel weerlegbaar zijn een verschillende status kunnen hebben, afhankelijk van de relaties tussen de zinnen in kwestie.

Het grootste deel van dit proefschrift is gewijd aan een grondige studie van deze logica's. In het bijzonder formuleren we een semantiek die zo eenvoudig mogelijk is en tonen daarbij aan dat heel wat mogelijke semantieken (eindige deterministische semantieken, Kripke semantieken, niet-normale Kripke semantieken en "neighbourhood" semantieken) niet geschikt zijn. We formuleren ook een volledige bewijstheorie en veralgemenen deze tot elke niet-deterministische gevolgrelatie. We bestuderen tenslotte de meest voorkomende paradoxen in verband met informele bewijsbaarheid en gaan na wat er gebeurt als we de onderliggende logica veranderen van klassieke logica naar CABAT.