A Formally Verified Abstract Account of Gödel's Incompleteness Theorems

Andrei Popescu[1] and Dmitriy Traytel[2]

[1] Department of Computer Science, Middlesex University London, UK
[2] Institute of Information Security, Department of Computer Science, ETH Zürich, Switzerland

Abstract. We present an abstract development of Gödel's incompleteness theorems, performed with the help of the Isabelle/HOL theorem prover. We analyze sufficient conditions for the theorems' applicability to a partially specified logic. In addition to the usual benefits of generality, our abstract perspective enables a comparison between alternative approaches from the literature. These include Rosser's variation of the first theorem, Jeroslow's variation of the second theorem, and the Świerczkowski–Paulson semantics-based approach. As part of our framework's validation, we upgrade Paulson's Isabelle proof to produce a mechanization of the second theorem that does not assume soundness in the standard model, and in fact does not rely on any notion of model or semantic interpretation.

1 Introduction

Gödel's incompleteness theorems [8, 11] are landmark results in mathematical logic. Both theorems refer to consistent logical theories that satisfy some assumptions, notably that of "containing enough arithmetic." The first incompleteness theorem ($\mathcal{IT}_1$) says that there are sentences that the theory cannot decide (i.e., neither prove nor disprove); the second theorem ($\mathcal{IT}_2$) says that the theory cannot prove (an internal formulation of) its own consistency. It is generally accepted that $\mathcal{IT}_1$ and $\mathcal{IT}_2$ have a wide scope, covering many logics and logical theories. However, when it comes to rigorous presentation, typically these results are only proved for particular, albeit paradigmatic cases, such as theories of arithmetic or hereditarily finite (HF) sets, within classical first-order logic (FOL); and even in these cases the constructions and proofs tend to be "incomplete and (apparently) irremediably messy" [3, p.16]. Hence, the theorems' scope remains largely unexplored on a rigorous/formal basis.

The emergence of powerful theorem provers has changed the rules of the game and, we argue, the expectation. Using interactive theorems provers, we can reliably keep track of all the constructions and their properties. Proof automation (often powered by fully automatic provers [16, 26]), makes complete, fully rigorous proofs feasible. And indeed, researchers have successfully met the challenge of mechanizing $\mathcal{IT}_1$ [13, 23, 25, 31] and recently $\mathcal{IT}_2$ [25]. Besides reassurance, these verification *tours de force* have brought superior technical insight into the theorems. But they have taken place within the same solitary confinement of scope as the informal proofs.

This paper takes steps towards a more comprehensive prover-backed exploration of the incompleteness theorems, based on a detailed analysis of their underlying assumptions. We use Isabelle/HOL [22] to establish general conditions under which these theorems apply to a partially specified logic. Our formalization is publicly available [28].

We start with a notion of logic (Section 2) whose terms, formulas and provability relation are kept abstract (Section 2.1). In particular, substitution and free variables are not defined, but axiomatized by some general properties. On top of this logic substratum, we consider an arithmetic substratum, consisting of a set of closed terms called *numerals* and an order-like relation (Section 2.2). Also factored in our abstract framework are encodings of formulas and proofs into numerals, the representability of various functions and relations as terms or formulas (Section 2.3), variations of the Hilbert-Bernays-Löb derivability conditions [14, 21] (Section 2.4), and standard models (Section 2.5).

Overall, our assumptions capture the notion of "containing enough arithmetics" in a general and flexible way. It is general because only few assumptions are made about the exact nature of formulas and numerals. It is flexible because different versions of the incompleteness theorems consider their own "amount of arithmetics" that makes it "enough," as proper subsets of these assumptions. Indeed, our formalization of the theorems (Section 3) proceeds in an austere-buffet style: Every result picks just enough infrastructure needed for it to hold—ranging from diagonalization which requires very little (Section 3.1) to Rosser's version of $\mathcal{IT}_1$ which is quite demanding. This approach caters for a sharp comparison between different formulations of the theorems, highlighting their trade-offs: Gödel's original formulation of $\mathcal{IT}_1$ versus Rosser's improvement (Section 3.2), proof-theoretic versus semantic versions of $\mathcal{IT}_1$ (Section 3.2), and Gödel's original formulation of the $\mathcal{IT}_2$ versus Jeroslow's improvement (Section 3.3).

Abstractness is our development's main strength, but also a potential weakness: Are our hypotheses reasonable? Are they consistent? These questions particularly concern our axiomatization of free variables and substitution—a notoriously error-prone area. As a remedy, we instantiate our framework to Paulson's semantics-based $\mathcal{IT}_1$ and $\mathcal{IT}_2$ for HF set theory [25], also performing an upgrade of Paulson's $\mathcal{IT}_2$ to a more general and more standard formulation: not restricted to sound theories, but applicable to any consistent theory (Section 4).

In the rest of this section, we discuss some guiding principles we followed when developing the formalization, and place our line of work in the context of related work.

Formal Design Principles Our long-term goal is a framework that makes it easy to instantiate the incompleteness theorems and related results to different logics. This is a daunting task, especially for $\mathcal{IT}_2$, where a lot of seemingly logic-specific technicalities are required to even formulate the theorem. The challenge is to push as much as possible of the technical constructions and lemmas to a largely logic-independent layer.

To this end, we strive to make minimal assumptions in terms of structure and properties when inferring the results—we will call this the *Economy* principle. For example, we do not define, but axiomatize syntax in terms of a minimal amount of structure. We assume a generic single-point substitution, then define simultaneous substitution and infer its properties. This is laborious, but worthwhile: Any logic that provides a single-point substitution satisfying our assumptions gets the simultaneous substitution for free.

As another instance of Economy, when faced with two different ways of formulating a theorem's conclusion we prefer the one that is *stronger under fewer assumptions*. (And dually, we prefer weakness for a theorem's assumptions.) For example, we discuss two variants of consistency: (1) "does not prove false" or (2) "there exists no formula such that itself and its negation are provable" (Section 3.3). The two statements are equivalent

at the meta-level, but the situation changes when they are represented as object-logic formulas: (1) implies (2) under mild assumptions, but the converse requires a lot more knowledge about the provability representation. So in our abstract theorems we strive to conclude (1) rather than (2). Even if (2) would also imply (1) in all reasonable instances, why postpone for the instantiation time any facts that we can acquire abstractly?

Applying the Economy principle not only stocks up generality for instantiations, but also accurately outlines trade-offs: How much does it cost (in terms of other added assumptions) to improve the conclusion, or to weaken an assumption of a theorem? For example, an Economy-based proof of Rosser's variant of $\mathcal{IT}_1$ reveals how much arithmetic we must factor in for weakening the ω-consistency assumption into consistency.

Related Work In his seminal publication, Gödel gave a proof of $\mathcal{IT}_1$ and the rough proof idea of $\mathcal{IT}_2$ [11]. Hilbert and Bernays gave a first detailed proof of $\mathcal{IT}_2$ [14]. Subsequently, a vast amount of literature was dedicated to the (re)formulation, proof, and analysis of these results [3,29,32,33]. The now canonical line of reasoning goes through the three derivability conditions devised by Bernays and Hilbert [14] and simplified by Löb [21]: One condition is used for $\mathcal{IT}_1$, and all three conditions for $\mathcal{IT}_2$. The derivability conditions have inspired a new branch of modal logic called provability logic [3]. Jeroslow has argued that one condition is redundant when proving $\mathcal{IT}_2$ [15].

Kreisel [18] and Jeroslow [15] were the first to study abstract conditions on logics under which the incompleteness theorems apply. Buldt [4] surveys the state of the art focusing on $\mathcal{IT}_1$, also sketching the applicability to non-standard logics. Our approach appears to be more abstract than all previous approaches: Being based on generic syntax and provability and truth predicates, it resembles the style of institution-independent model theory [7, 12] and our previous work on abstract completeness [2] and completeness of ordered resolution [30]. However, there are certainly dimensions of generality and strength of this vast topic that our formalized work does not (yet) explore. These include quantifier-free logics [15] and arithmetical hierarchy refinements [17]. Our syntax axiomatization is inspired by algebraic theories of the λ-calculi syntax [9, 10, 27].

In the realm of mechanical proofs, $\mathcal{IT}_1$ has been proved by Shankar [31] in the Boyer-Moore prover, Harrison in HOL Light [13] and O'Connor in Coq [23]. $\mathcal{IT}_2$ has only been proved recently—by Paulson in Isabelle/HOL [24,25] (who also proved $\mathcal{IT}_1$). All these mechanizations target theories over a fixed language in classical FOL: that of arithmetic (Harrison and O'Connor) and that of HF sets (Shankar and Paulson). These mechanizations are mostly focused on "getting all the work done" in a particular setting (although Harrison targets a more abstract class of theories in the given language). On their way to $\mathcal{IT}_1$, Shankar and O'Connor also prove the representability of all partial, respectively primitive recursive functions—which are important standalone results.

By contrast, we explore conditions that enable different formulations for an abstract logic, where aspects such as recursiveness are below our abstraction level. The two approaches are complementary, and they both contribute to formally taming the complex ramifications of the incompleteness theorems. When instantiating our abstract assumptions to recover and upgrade Paulson's results, we took advantage of Paulson's substantial work on proving the many low-level lemmas towards the derivability conditions. More should be done at an abstract level to avoid duplicating some of these laborious lemmas when instantiating the theorems to different logics. This will be future work.

2 Abstract Assumptions

Roughly, the incompleteness theorems are considered to hold for logical theories that (1) contain enough arithmetic and (2) are "effective" in that they themselves can be arithmetized. Our goal is to give a general expression of these favorable conditions. To this end, we identify some logic and arithmetic substrata consisting of structure and axioms that express the containment of (various degrees of) arithmetic more abstractly and flexibly than relative interpretations [35]. We also identify abstract notions of encodings and representability that have just what it takes for a working arithmetization.

2.1 The logical substratum

We start with some unspecified sets of variables (Var, ranged over by x, y, z), terms (Term, ranged over by s, t) and formulas (Fmla, ranged over by φ, ψ, χ). We assume that variables are particular terms, $\mathsf{Var} \subseteq \mathsf{Term}$, and that Var is infinite. Free-variables and substitution operators, FVars and $\_\,[\_/\_]$, are assumed for both terms and formulas. We think of $\mathsf{FVars}(t)$ as the (finite) set of free variables of the term t, and similarly for formulas. We call *sentence* any formula with no free variable, and let Sen denote the set of sentences. We think of $s\,[t/x]$ as the term obtained from s by the (capture-avoiding) substitution of t for the free occurrences of variable x; we think of $\varphi\,[t/x]$ as the formula obtained from φ by the substitution of t for the free occurrences of variable x.

In FOL, terms introduce no bindings, so any occurring variable is free. FOL terms fall under our framework, and so do terms with bindings as in λ-calculi and higher-order logic (HOL). To achieve this degree of inclusiveness while also being able to prove interesting results, we work under some well-behavedness assumptions about the free-variables and substitution operators. For example, free-variables distribute over substitution, $\mathsf{FVars}\,(\varphi\,[s/x]) = \mathsf{FVars}(\varphi) - \{x\} \cup \mathsf{FVars}(s)$ if $x \in \mathsf{FVars}(\varphi)$, and substitution is compositional, $\varphi\,[s_1/x_1]\,[s_2/x_2] = \varphi\,[s_2/x_2]\,[(s_1\,[s_2/x_2])\,/\,x_1]$ if $x_1 \neq x_2$ and $x_1 \notin \mathsf{FVars}(s_2)$. The full list of our generic syntax axioms is shown in Appendix A.1.

The incompleteness theorems rely heavily on simultaneous substitution, written $\varphi\,[t_1/x_1, \ldots, t_n/x_n]$, whose properties are tricky to formalize—for example, Paulson's formalization paper dedicates them ample space [25, 6.2]. To address this problem once and for all generically, we define simultaneous substitution from the single-point substitution, $\varphi\,[t/x]$, and infer its properties from the single-point substitution axioms. For example, we prove that $\mathsf{FVars}\,(\varphi\,[s_1/x_1, \ldots, s_n/x_n]) = \mathsf{FVars}(\varphi) \cup \bigcup \{\mathsf{FVars}(s_i) - \{x_i\} \mid i \in \{1, \ldots, n\} \text{ and } x_i \in \mathsf{FVars}(\varphi)\}$. The technicalities are delicate: To avoid undesired variable replacements, $\varphi\,[s_1/x_1, \ldots, s_n/x_n]$ must be defined as $\varphi\,[y_1/x_1] \ldots [y_n/x_n]\,[s_1/y_1] \ldots [s_n/y_n]$ for some fresh $y_1, \ldots, y_n$, the choice of which we must show to be immaterial. This definition's complexity is reflected in the properties' proofs. But again, this one-time effort benefits any "customer" logic: In exchange for a well-behaved single-point substitution, it gets back a well-behaved simultaneous substitution.

We let $v_1, v_2, \ldots$ be fixed mutually distinct variables. We write Fmla_k for the set of formulas whose free variables are precisely $\{v_1, \ldots, v_k\}$, and $\mathsf{Fmla}_k^{\subseteq}$ for the set of formulas whose variables are among $\{v_1, \ldots, v_k\}$. Note that $\mathsf{Fmla}_k \subseteq \mathsf{Fmla}_k^{\subseteq}$ and $\mathsf{Fmla}_0 = \mathsf{Fmla}_0^{\subseteq} = \mathsf{Sen}$. Given $\varphi \in \mathsf{Fmla}_k^{\subseteq}$, we write $\varphi\,(t_1, \ldots, t_n)$ instead of $\varphi\,[t_1/v_1, \ldots, t_n/v_n]$.

In addition to free variables and substitution, our theorems will require formulas to be equipped with term equality ($\equiv$), Boolean connectives ($\bot$, $\top$, $\to$, $\neg$, $\wedge$, $\vee$), universal and existential quantifiers ($\forall$, $\exists$). When needed, we will assume them not as constructors (syntax builders), but as operators on terms and formulas, e.g., $\equiv\ : \mathsf{Term} \to \mathsf{Term} \to \mathsf{Fmla}$, $\bot \in \mathsf{Fmla}$, $\forall\ : \mathsf{Var} \times \mathsf{Fmla} \to \mathsf{Fmla}$. This caters for logics that do not have them as primitives, but can define them. For example, higher-order logic (HOL) defines all connectives and quantifiers from λ-abstraction and either equality or implication.

We fix a unary relation $\vdash \subseteq \mathsf{Fmla}$ on formulas, called *provability*. We write $\vdash \varphi$ instead of $\varphi \in \vdash$, and say the formula φ is *provable*. Whenever certain formula connectives or quantifiers are assumed present, we will assume that $\vdash$ behaves intuitionistically w.r.t. them—namely, we assume the usual (Hilbert-style) intuitionistic FOL axioms with respect to the abstract connectives and quantifiers. Stronger systems, such as those of classical logic, also satisfy these assumptions.

Consistency, denoted Con, is defined as the impossibility to prove false, namely $\nvdash \bot$. Another central concept is ω-consistency—we carefully choose a formulation that works intuitionistically, with conclusion reminiscent of Gödel's negative translation [6]:

OCon: For all $\varphi \in \mathsf{Fmla}_1^{\subseteq}$, if $\vdash \neg\, \varphi(n)$ for all $n \in \mathsf{Num}$ then $\nvdash \neg\, \neg (\exists x.\, \varphi(x))$.

Assuming classic deduction in $\vdash$, this is equivalent to the standard formulation: For all $\varphi \in \mathsf{Fmla}_1^{\subseteq}$, it is not the case that $\vdash \varphi(n)$ for all $n \in \mathsf{Num}$ and $\vdash \neg\, (\forall x.\, \varphi(x))$.

Occasionally we will need to consider not only provability, but also explicit proofs. In such cases, we fix a set Proof of (entities which we call) *proofs*, ranged over by p, q, and a binary relation between proofs p and sentences φ, written $p \Vdash \varphi$, of which we think of as stating that p is a proof of φ. We also assume that $\vdash$ and $\Vdash$ are related as expected, in that provability means the existence of a proof:

$\mathsf{Rel}_{\vdash}^{\Vdash}$: For all $\varphi \in \mathsf{Sen}$, $\vdash \varphi$ iff there exists $p \in \mathsf{Proof}$ such that $p \Vdash \varphi$.

2.2 The arithmetic substratum

We extend the generic syntax assumptions with a subset $\mathsf{Num} \subseteq \mathsf{Term}$, of *numerals*, ranged over by m, n, which are assumed to be closed, i.e., have no free variables.

Convention 1. In all the shown results we implicitly assume: (1) the generic syntax (free variable and substitution) axioms, (2) at least $\to$ and $\bot$ plus whatever connectives and quantifiers appear in the statement, (3) closedness of $\vdash$ under intuitionistic deduction rules, and (4) the existence of numerals. Other assumptions (e.g., order-like relation axioms, consistency, standard models, etc.) will be indicated explicitly.

On one occasion, we will assume an order-like binary relation modeled by a formula $\prec \in \mathsf{Fmla}_2$. We write $t_1 \prec t_2$ instead of $\prec (t_1, t_2)$ and $\forall x \prec n.\, \varphi$ instead of $\forall x.\, x \prec n \to \varphi$. It turns out that at our level of abstraction it does not matter whether $\prec$ is a strict or a non-strict order. Indeed, we only require the following two properties, where $x \in M$ denotes $\bigvee_{m \in M} x \equiv m$ and $\bigvee$ expresses the disjunction of a finite set of formulas:

Ord_1: For all $\varphi \in \mathsf{Fmla}_1$ and $n \in \mathsf{Num}$, if $\vdash \varphi(m)$ for all $m \in \mathsf{Num}$, then $\vdash \forall x \prec n.\, \varphi(x)$.
Ord_2: For all $n \in \mathsf{Num}$, there exists a finite set $M \subseteq \mathsf{Num}$ such that $\vdash \forall x.\, x \in M \vee n \prec x$.

Ord_1 states that if a property φ is provable for all numerals, then its universal quantification bounded by any given numeral n is also provable. Having in mind the arith-

metic interpretation of numerals, it would also make sense to assume a stronger version of Ord_1, replacing "if $\vdash \varphi(m)$ for all $m \in \mathsf{Num}$" by the weaker hypothesis "if $\vdash \varphi(m)$ for all $m \in \mathsf{Num}$ such that $\vdash m \prec n$". But this stronger version will not be needed.

Ord_2 states that, for any numeral n, any element x in the domain of discourse is either greater than n or equal to one of a finite set M of numerals. If we instantiate our syntax to that of first-order arithmetic, then the natural number model satisfies Ord_1 and Ord_2 when interpreting $\prec$ as either $<$ or $\leq$. Moreover, these properties are provable in intuitionistic Robinson arithmetic, again for both $<$ and $\leq$.

2.3 Encodings and representability

Central in the incompleteness theorems are functions that encode formulas and proofs as numerals, $\langle \_ \rangle : \mathsf{Fmla} \to \mathsf{Num}$ and $\langle \_ \rangle : \mathsf{Proof} \to \mathsf{Num}$. For our abstract results, the encodings are not required to be injective or surjective.

Let $A_1, \ldots, A_m$ be sets, and let, for each of them, $\langle \_ \rangle : A_i \to \mathsf{Num}$ be an "encoding" function to numerals. Then, an m-ary relation $R \subseteq A_1 \times \ldots \times A_m$ is said to be *represented* by a formula $\textcircled{R} \in \mathsf{Fmla}_m$ if the following hold for all $(a_1, \ldots, a_m) \in A_1 \times \ldots \times A_m$:

- $(a_1, \ldots, a_m) \in R$ implies $\vdash \textcircled{R}(\langle a_1 \rangle, \ldots, \langle a_m \rangle)$
- $(a_1, \ldots, a_m) \notin R$ implies $\vdash \neg\textcircled{R}(\langle a_1 \rangle, \ldots, \langle a_m \rangle)$

Let A be another set with $\langle \_ \rangle : A \to \mathsf{Num}$. An m-ary function $f : A_1 \times \ldots A_m \to A$ is said to be *represented* by a formula $\textcircled{f} \in \mathsf{Fmla}_{m+1}$ if for all $(a_1, \ldots, a_m) \in A_1 \times \ldots \times A_m$:

- $\vdash \textcircled{f}(\langle a_1 \rangle, \ldots, \langle a_m \rangle, \langle f\,(a_1, \ldots, a_m) \rangle)$
- $\vdash \forall x, y.\ \textcircled{f}(\langle a_1 \rangle, \ldots, \langle a_m \rangle, x) \wedge \textcircled{f}(\langle a_1 \rangle, \ldots, \langle a_m \rangle, y) \to x = y$

The notion of a function being represented is stronger than that of its graph being represented (as a relation)—but with enough deductive power they are equivalent [32, §16]. We will need an even stronger notion: A function f as above is *term-represented* by an operator $\textcircled{f} : \mathsf{Term}^m \to \mathsf{Term}$ if $\vdash \textcircled{f}(\langle a_1 \rangle, \ldots, \langle a_m \rangle) \equiv \langle f\,(a_1, \ldots, a_m) \rangle$ for all $(a_1, \ldots, a_m) \in A_1 \times \ldots \times A_m$. When the formula by which a relation/function P is represented or term-represented is irrelevant, we call P *representable* or *term-representable*.

We will also need an enhancement of relation representability: Given $i < m$, we call the representation of an m-ary relation R by $\textcircled{R}$ *i-clean* if $\vdash \neg\,\textcircled{R}(n_1, \ldots, n_m)$ for all $n_1, \ldots, n_m$ such that n_i is outside the image of $\langle \_ \rangle$ (i.e., there is no $a \in A_i$ with $n_i = \langle a \rangle$). Cleanness would be trivially satisfied if the encodings were surjective. However, surjectivity is not a reasonable assumption. For example, most of the numeric encodings used in the literature are injective but not surjective.

We let $\mathsf{S} : \mathsf{Fmla}_1 \to \mathsf{Sen}$ be the *self-substitution* function, which sends any $\varphi \in \mathsf{Fmla}_1$ to $\varphi(\langle \varphi \rangle)$, i.e., to the sentence obtained from φ by substituting the encoding of φ for the unique variable of φ. An alternative is the following "soft" version of S, which sends any $\varphi \in \mathsf{Fmla}_1$ to $\exists v_1.\ v_1 \equiv \langle \varphi \rangle \wedge \varphi$, where v_1 is the single free variable of φ. The soft version yields provably equivalent formulas and has the advantage that it is easier to represent inside the logic, since it does not require formalizing the complexities of capture-avoiding substitution. All our results involving S have been proved for both versions.

We will consider the properties $\mathsf{Repr}_{\neg}$, $\mathsf{Repr}_{\mathsf{S}}$, and $\mathsf{Repr}_{\Vdash}$, stating the representability of the functions $\neg$ and S, and of the relation $\Vdash$; and $\mathsf{TRepr}_{\neg}$, stating the term-representability of $\neg$. In addition, $\mathsf{Clean}_{\Vdash}$ will state that the considered representation of $\Vdash$ is 1-clean, i.e., it is clean on the proof component. For the representing formulas for the above relations and functions we will use their circled names, $\textcircled{\neg}$, $\textcircled{\Vdash}$, etc.; for example, $\mathsf{Repr}_{\Vdash}$ means that (1) $p \Vdash \varphi$ implies $\vdash \textcircled{\Vdash}(\langle p \rangle, \langle \varphi \rangle)$ and (2) $p \nVdash \varphi$ implies $\vdash \neg\, \textcircled{\Vdash}(\langle p \rangle, \langle \varphi \rangle)$ for all $p \in \mathsf{Proof}$ and $\varphi \in \mathsf{Sen}$. $\vdash \boxed{\neg}\, \langle \varphi \rangle \equiv \langle \neg\, \varphi \rangle$ for all $\varphi \in \mathsf{Sen}$.

2.4 Derivability conditions

Most of our assumptions refer to representability. An important exception is the provability relation $\vdash$, for which only a weakening of representability is reasonable. Let $\textcircled{\vdash} \in \mathsf{Fmla}_1$ be the formula for this task. We consider the following assumptions about $\textcircled{\vdash}$, known as the Hilbert-Bernays-Löb derivability conditions:

HBL_1: $\vdash \varphi$ implies $\vdash \textcircled{\vdash}\langle \varphi \rangle$ for all $\varphi \in \mathsf{Sen}$.
HBL_2: $\vdash \textcircled{\vdash}\langle \varphi \rangle \wedge \textcircled{\vdash}\langle \varphi \to \psi \rangle \to \textcircled{\vdash}\langle \psi \rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.
HBL_3: $\vdash \textcircled{\vdash}\langle \varphi \rangle \to \textcircled{\vdash}\langle \textcircled{\vdash}\langle \varphi \rangle \rangle$ for all $\varphi \in \mathsf{Sen}$.

Above and elsewhere, to lighten notation we omit parentheses when instantiating one-variable formulas with encodings of formulas—e.g., writing $\textcircled{\vdash}\langle \varphi \rangle$ instead of $\textcircled{\vdash}(\langle \varphi \rangle)$.

HBL_1 states that, if a sentence is provable, then its encoding is also provable inside the representation. HBL_3 is roughly a formulation of HBL_1 "one level up," inside the proof system $\vdash$. Finally, note that the provability relation is closed under *modus ponens*, in that $\vdash \varphi$ and $\vdash \varphi \to \psi$ implies $\vdash \psi$ for all $\varphi, \psi \in \mathsf{Sen}$. Thus, HBL_2 roughly states the same property inside the proof system. In short, the derivability conditions state that the representation of provability acts partly similarly to the provability relation. Note that the representability of "proof of" implies HBL_1, taking $\textcircled{\vdash}(x)$ to be $\exists y.\, \textcircled{\Vdash}(y, x)$.

Convention 2. In this paper, we focus on the standard provability representation: Whenever we assume explicit proofs and representability of "proof of," the formula $\textcircled{\vdash}$ will be defined from $\textcircled{\Vdash}$ as shown above.

We will also be interested in the following variations of the derivability conditions:

HBL_4: $\vdash \textcircled{\vdash}\langle \varphi \rangle \wedge \textcircled{\vdash}\langle \psi \rangle \to \textcircled{\vdash}\langle \varphi \wedge \psi \rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.
$\mathsf{HBL}_1^{\Leftarrow}$: $\vdash \textcircled{\vdash}\langle \varphi \rangle$ implies $\vdash \varphi$ for all $\varphi \in \mathsf{Sen}$.
SHBL_3: $\vdash \textcircled{\vdash}(t) \to \textcircled{\vdash}\langle \textcircled{\vdash}(t) \rangle$ for all closed terms t.
WHBL_2: $\vdash \varphi \to \psi$ implies $\vdash \textcircled{\vdash}\langle \varphi \rangle \to \textcircled{\vdash}\langle \psi \rangle$ for all $\varphi, \psi \in \mathsf{Sen}$.

HBL_4 has a similar flavor as HBL_2, but refers to conjunction: It states that the conjunction introduction rule holds inside the proof system. $\mathsf{HBL}_1^{\Leftarrow}$ is the converse of HBL_1. Finally, SHBL_3 is a strengthening of HBL_3 holding for all closed terms and not only those that encode sentences, and (if we assume HBL_1) WHBL_2 is a weakening of HBL_2.

2.5 Standard models

We fix a unary relation $\models \subseteq \mathsf{Sen}$, representing *truth of a sentence in the standard model*. We write $\models \varphi$ instead of $\varphi \in \models$, and read it as "φ is true." We consider the assumptions:

$\mathsf{Syn}_{\models}$: Syntactic entities (logical connectives and quantifiers) handle truth as expected:
(1) $\not\models \bot$; (2) for all $\varphi, \psi \in \mathsf{Sen}$, $\models \varphi$ and $\models \varphi \to \psi$ imply $\models \psi$;

(3) for all $\varphi \in \mathsf{Fmla}_1$, if $\models \varphi(n)$ for all $n \in \mathsf{Num}$ then $\models \forall x.\, \varphi(x)$;
(4) for all $\varphi \in \mathsf{Fmla}_1$, if $\models \exists x.\, \varphi(x)$ then $\models \varphi(n)$ for some $n \in \mathsf{Num}$;
(5) for all $\varphi \in \mathsf{Sen}$, $\models \varphi$ or $\models \neg\, \varphi$.
Soundness (of provability with respect to truth): $\vdash \varphi$ implies $\models \varphi$ for all $\varphi \in \mathsf{Sen}$.

$\mathsf{Syn}_{\models}$(1–4) only contains a partial description of the syntactic entities' behavior—corresponding to elimination rules for $\bot$, $\to$ and $\exists$ and introduction rule for $\forall$. For our results this suffices. $\mathsf{Syn}_{\models}$(5) states that standard models decide every sentence.

On his way to formalizing $\mathcal{IT}_2$ for extensions of the HF set theory, after proving HBL_1 Paulson notes [25, p.21]: "The reverse implication [namely $\mathsf{HBL}_1^{\Leftarrow}$], despite its usefulness, is not always proved." In his abstract account, Buldt also assumes $\mathsf{HBL}_1^{\Leftarrow}$ in his most general formulation of $\mathcal{IT}_1$ [4, Theorem 3.1]; that formulation has in mind not necessarily the standard provability representation (our Convention 2), but any formula that weakly represents $\vdash$, which is acceptable for $\mathcal{IT}_1$ but not for $\mathcal{IT}_2$ [1].

We avoid such $\mathcal{IT}_1$ versus $\mathcal{IT}_2$ divergence by remaining focused on the standard provability representation. In this case, for arithmetics and related theories, $\mathsf{HBL}_1^{\Leftarrow}$ cannot be inferred without assuming soundness in the standard model (which Paulson does), or at least ω-consistency. We can depict the situation abstractly, without knowing how standard models look like:

Lemma 3. (1) Assume $\mathsf{Rel}_{\vdash}^{\Vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ and OCon. Then $\mathsf{HBL}_1^{\Leftarrow}$ holds.
(2) Assume $\mathsf{Soundness}$ and $\mathsf{Syn}_{\models}$(1,2,3). Then OCon holds.
(3) Assume $\mathsf{Rel}_{\vdash}^{\Vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$, $\mathsf{Soundness}$ and $\mathsf{Syn}_{\models}$(1,2,4). Then $\models \textcircled{\scriptsize P}\langle\varphi\rangle$ implies $\vdash \varphi$ for all $\varphi \in \mathsf{Sen}$. In particular, $\mathsf{HBL}_1^{\Leftarrow}$ holds.

Thus, staying in a proof-theoretic world, ω-consistency ensures $\mathsf{HBL}_1^{\Leftarrow}$ if the "proof of" relation is cleanly represented (1). In turn, ω-consistency is ensured by minimal semantic requirements, including the soundness of provability (2). Finally, putting together representability and semantics, we can infer something stronger than $\mathsf{HBL}_1^{\Leftarrow}$: That the mere truth (and not just the provability) of a sentence's provability representation implies the provability of the sentence itself (3).

It follows from either points (1,2) or point (3) of the lemma that, in the presence of standard models and soundness, clean representability of the "proof of" relation implies $\mathsf{HBL}_1^{\Leftarrow}$; and recall that it also implies HBL_1. So it implies an "iff" version of HBL_1: $\vdash \varphi$ if an only if $\vdash \textcircled{\scriptsize P}\langle\varphi\rangle$. Interestingly, a converse of this implication also holds. To state it, we initially assume there is no "outer" notion of proof (i.e., no set Proof and no relation $\Vdash$), but only an "inner" one, given by a formula $\mathsf{P} \in \mathsf{Fmla}_2$ such that:

$\mathsf{Rel}^{\mathsf{P}}_{\textcircled{\scriptsize P}}$: $\vdash \textcircled{\scriptsize P}\langle\varphi\rangle \;\leftrightarrow\; \exists x.\, \mathsf{P}(x, \langle\varphi\rangle)$.
$\mathsf{Compl}_{\mathsf{P}}$: $\models \mathsf{P}(n, \langle\varphi\rangle)$ implies $\vdash \mathsf{P}(n, \langle\varphi\rangle)$ for all $n \in \mathsf{Num}$ and $\varphi \in \mathsf{Sen}$.
$\mathsf{Compl}_{\neg\mathsf{P}}$: $\models \neg\, \mathsf{P}(n, \langle\varphi\rangle)$ implies $\vdash \neg\, \mathsf{P}(n, \langle\varphi\rangle)$ for all $n \in \mathsf{Num}$ and $\varphi \in \mathsf{Sen}$.

$\mathsf{Rel}^{\mathsf{P}}_{\textcircled{\scriptsize P}}$ is the inner version of $\mathsf{Rel}_{\vdash}^{\Vdash}$: It expresses that, *inside the representation*, proofs and provability are connected as expected. $\mathsf{Compl}_{\mathsf{P}}$ and $\mathsf{Compl}_{\neg\mathsf{P}}$ state that provability is complete on P statements about formula encodings, as well as their negations; in traditional settings, this is true thanks to P being a bounded arithmetical formula (Δ_0). Now the converse result states that, thanks to (standard models and) the "iff" version of HBL_1, we can define an outer notion of proof that is represented by the inner notion P:

Lemma 4. Assume $\mathsf{Rel}^{\mathsf{P}}_{\textcircled{\vdash}}$, Soundness, $\mathsf{Syn}_{\models}$(4,5), $\mathsf{Compl}_{\mathsf{P}}$, $\mathsf{Compl}_{\neg\mathsf{P}}$, HBL_1 and $\mathsf{HBL}^{\Leftarrow}_1$. Take $\mathsf{Proof} = \mathsf{Num}$ and define $\Vdash$ by $n \Vdash \varphi$ iff $\vdash \mathsf{P}\,(n, \langle\varphi\rangle)$. Then $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$ hold, with $\Vdash$ being represented by P.

3 Abstract Incompleteness Theorems

After last section's preparations, we are now ready to discuss different versions of the incompleteness theorems and their major lemmas, based on alternative assumptions.

3.1 Diagonalization

The formula diagonalization technique (due to Gödel and Carnap [5]) yields "self-referential" sentences. All we need for it to work is the representability of substitution.

Prop 5. Assuming $\mathsf{Repr}_{\mathsf{S}}$, for all $\psi \in \mathsf{Fmla}_1$ there exists $\varphi \in \mathsf{Fmla}_1$ with $\vdash \varphi \leftrightarrow \psi\langle\varphi\rangle$.

A sentence $\varphi \in \mathsf{Sen}$ is called a *Gödel sentence* if $\vdash \varphi \leftrightarrow \neg\, \textcircled{\vdash}\langle\varphi\rangle$; it is called a *Rosser sentence* if $\vdash \varphi \leftrightarrow \neg\,(\exists x.\, \textcircled{\vdash}(x, \langle\varphi\rangle) \wedge \mathsf{RosserTwist}(x, \langle\varphi\rangle))$, where we define $\mathsf{RosserTwist}(x, y) = \forall x'.\, x' \prec x \to \forall y'.\, \textcircled{\neg}(y, y') \to \neg\, \textcircled{\vdash}(x', y')$. The existence of Gödel and Rosser sentences follows immediately from diagonalization.

Prop 6. Assuming $\mathsf{Repr}_{\mathsf{S}}$, there exist Gödel and Rosser sentences.

Thus, any Gödel sentence is provably equivalent to the negation of its own provability; in Gödel's words [11], it "says about itself that it is not provable." A Rosser sentence φ asserts its own unprovabilty in a weaker fashion: Rather than saying "Myself, φ, am not provable" (i.e., "it is not the case that there exists a proof p of φ"), it says "it is not the case that there exists a proof p of φ such that, for all smaller proofs q, q is not a proof of $\neg\,\varphi$." Here, "smaller" refers to the order the encoding of proofs as numerals imposes.

3.2 The incompleteness theorems

$\mathcal{IT}_1$ identifies sentences that are neither provable nor disprovable—which often holds for Gödel and Rosser sentences with the help of a provability relation satisfying HBL_1.

Prop 7. Assume Con and HBL_1. Then $\nvdash \mathsf{G}$ for all Gödel sentences G.

For showing that the Gödel sentences are not disprovable, a standard route is to assume explicit proofs, strengthen the consistency assumption to ω-consistency, and strengthen HBL_1 to representability of the "proof of" relation.

Prop 8. Assume OCon, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$. Then $\nvdash \neg\, \mathsf{G}$ for all Gödel sentences G.

Proof. Let G be a Gödel sentence. We prove $\nvdash \neg\, \mathsf{G}$ by contradiction. Assume (1) $\vdash \neg\, \mathsf{G}$.
- By consistency (which is implied by OCon), we obtain $\nvdash \mathsf{G}$.
- From this and $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $p \nVdash \mathsf{G}$ for all $p \in \mathsf{Proof}$.
- From this, $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$, we obtain $\vdash \neg\, \textcircled{\vdash}(n, \langle\mathsf{G}\rangle)$ for all $n \in \mathsf{Num}$.
- From this and OCon, we obtain $\nvdash \neg\,\neg\,\exists x.\, \textcircled{\vdash}(x, \langle\mathsf{G}\rangle)$, i.e., $\nvdash \neg\,\neg\, \textcircled{\vdash}\langle\mathsf{G}\rangle$.
- Hence, since G is a Gödel sentence, we obtain $\nvdash \neg\, \mathsf{G}$, which contradicts (1). □

While the line of reasoning in the above proof is mostly well-known, it contains two subtle points about which the literature is not explicit (due to the usual focus on classical first-order arithmetic and particular choices of encodings).

First, we must assume the representation of the "proof of" relation to be 1-*clean*, i.e., clean with respect to the proof component. Indeed, the argument crucially relies on converting the statement "$p \nVdash \mathsf{G}$ for all $p \in \mathsf{Proof}$" into "$\vdash \neg \textcircled{\scriptsize P}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$," which is only possible for 1-clean encodings. This assumption will be repeatedly needed in later results. By contrast, cleanness is never required with respect to the sentence component of "proof of" or for the provability relation (which only involves sentence encodings). In short, cleanness is only needed for proofs, not for sentences.

Second, to reach the desired contradiction for our intuitionistic proof system $\vdash$, from "$\vdash \neg \textcircled{\scriptsize P}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$" it is not sufficient to employ standard ω-consistency, which would only give us $\nvdash \exists x.\, \textcircled{\scriptsize P}(x, \langle \mathsf{G} \rangle)$, i.e., $\nvdash \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$; the last together with $\vdash \mathsf{G} \leftrightarrow \neg \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$ would be insufficient for obtaining $\nvdash \neg \mathsf{G}$. However, our stronger version of ω-consistency, OCon, does the trick. $\mathcal{IT}_1$ now follows by putting together Props. 6–8:

Theorem 9. ($\mathcal{IT}_1$) Assume OCon, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ and $\mathsf{Repr}_{\mathsf{S}}$. Then:

(1) There exists a Gödel sentence. (2) $\nvdash \mathsf{G}$ and $\nvdash \neg \mathsf{G}$ for all Gödel sentences G.

Rosser's contribution to $\mathcal{IT}_1$ was an ingenious trick for weakening the ω-consistency assumption into plain consistency—as such, it is usually seen as a *strict improvement* over Gödel's version. While this is true for the concrete case of FOL theories extending arithmetic, from an abstract perspective the situation is more nuanced: The improvement is achieved at the cost of asking more from the logic. Our framework makes this trade-off clearly visible. The idea is to use Rosser sentences instead of Gödel sentences to "repair" the ω-consistency assumption of Theorem 9 (inherited from Prop. 8):

Theorem 10. (Rosser's $\mathcal{IT}_1$) Assume Con, Ord_1, Ord_2, $\mathsf{Repr}_{\neg}$, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ and $\mathsf{Repr}_{\mathsf{S}}$. Then:

(1) There exists a Rosser sentence. (2) $\nvdash \mathsf{R}$ and $\nvdash \neg \mathsf{R}$ for all Rosser sentences R.

Highlighted is the assumption trade-off between the two versions: Rosser's weakening of ω-consistency into consistency is paid by additionally assuming representability of negation and an order-like relation satisfying Ord_1 and Ord_2. Certainly, negation representability is not a big price, since for concrete logics this tends to be a lemma that is anyway needed when proving HBL_1. On the other hand, the ordering assumptions seem to be a significant generality gap in favor of Gödel's version. A clear manifestation of this gap is in our inference of a semantic version of $\mathcal{IT}_1$—which we obtain from Theorem 9 with the help of Lemmas 3(2) and 4:

Theorem 11. (Semantic $\mathcal{IT}_1$) Assume $\mathsf{Rel}^{\mathsf{P}}_{\textcircled{\scriptsize P}}$, $\mathsf{Soundness}$, $\mathsf{Syn}_{\models}$, $\mathsf{Compl}_{\mathsf{P}}$, HBL_1, $\mathsf{HBL}_1^{\Leftarrow}$, and $\mathsf{Repr}_{\mathsf{S}}$. Then:

(1) There exists a Gödel sentence. (2) $\models \mathsf{G}$, $\nvdash \mathsf{G}$ and $\nvdash \neg \mathsf{G}$ for all Gödel sentences G.

We have highlighted the assumptions specific to the semantic treatment. They replace OCon, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$ from the proof-theoretic Theorem 9. Also highlighted is the additional fact concluded: that the Gödel sentences are true.

We have inferred the semantic version from Gödel's proof-theoretic version (Theorem 9), and not from Rosser's variation (Theorem 10). This is because in the semantic version ω-consistency comes for free (from Lemma 3(2)). By contrast, for deploying Rosser's version we would need to explicitly consider the order-like relation with its own hypotheses. This would have led to a *strictly less general* abstract result (if we ignore the difference in the way Gödel and Rosser sentences are actually defined).

The semantic $\mathcal{IT}_1$ relies on $\mathsf{HBL}_1^{\Leftarrow}$. If we additionally commit to classical logic (i.e., assume $\vdash \neg\neg\varphi \to \varphi$), we can more directly show, taking advantage of $\mathsf{HBL}_1^{\Leftarrow}$, that the Gödel sentences are not disprovable, which immediately proves $\mathcal{IT}_1$:

Theorem 12. (Classical $\mathcal{IT}_1$) Assume classical logic, Con, HBL_1, $\mathsf{HBL}_1^{\Leftarrow}$, $\mathsf{Repr}_{\mathsf{S}}$. Then:
(1) There exists a Gödel sentence. (2) $\nvdash \mathsf{G}$ and $\nvdash \neg\,\mathsf{G}$ for all Gödel sentences G.

Even though $\mathcal{IT}_1$ uses a predicate $\textcircled{\vdash}$ that weakly represents provability, its conclusion (the existence of undecided sentences) is meaningful regardless of whether $\textcircled{\vdash}$ *adequately expresses provability*. By contrast, the meaning of $\mathcal{IT}_2$'s conclusion, the theory cannot prove its own consistency, relies on this (non-mathematical) "intensional" assumption [1]. In this case, consistency is adequately expressed by the sentence $\neg\,\textcircled{\vdash}\langle\bot\rangle$. The standard formulation (and proof) of $\mathcal{IT}_2$ uses all three derivability conditions:

Theorem 13. ($\mathcal{IT}_2$) Assume Con, HBL_1, HBL_2, HBL_3 and $\mathsf{Repr}_{\mathsf{S}}$. Then $\nvdash \neg\,\textcircled{\vdash}\langle\bot\rangle$.

3.3 Jeroslow's approach

Next we look into an alternative line of reasoning due to Jeroslow [15], often cited as a simplification of the canonical route to prove $\mathcal{IT}_2$ [29, 32, 33]. To study its features and pitfalls, we need some standard notation used by Jeroslow. A *pseudo-term* is a formula $\varphi \in \mathsf{Fmla}_m$ expressing a provably functional relation via "exists unique": $\vdash \forall x_1, \ldots, x_m.\ \exists! y.\ \varphi(x_1, \ldots, x_m, y)$. Next we only discuss the case $m = 2$; the general case is similar.

Notation 14. Given a pseudo-term $\varphi \in \mathsf{Fmla}_2$, we treat it as if it is a one-variable term:
- for any terms s and t, we write $t \equiv \varphi(s)$ instead of $\varphi(s, t)$;
- for any term s and formula $\psi \in \mathsf{Fmla}_1$, we write $\psi(\varphi(s))$ instead of $\exists y.\ \varphi(s, y) \wedge \psi(y)$.

This notation smoothly integrates pseudo-terms with terms: If $\vdash t \equiv \varphi(s)$ and $\vdash \psi(\varphi(s))$ then $\vdash \psi(t)$, where $\psi(t)$ denotes actual substitution of terms in formulas.

Jeroslow relies on an abstract class of m-ary functions, $\mathcal{F}_m \subseteq \mathsf{Num}^m \to \mathsf{Num}$, for all arities $m \in \mathbb{N}$, on which he considers following assumptions:

$\mathsf{Repr}_{\mathcal{F}}$: Every $f \in \mathcal{F}_m$ is represented by some pseudo-term $\textcircled{f} \in \mathsf{Fmla}_{m+1}$ under the identity encoding $\mathsf{Num} \to \mathsf{Num}$.
CapN: Some $\mathsf{N} \in \mathcal{F}_1$ correctly captures negation: $\mathsf{N}\langle\varphi\rangle = \langle\neg\,\varphi\rangle$ for all $\varphi \in \mathsf{Sen}$.
CapSS: Some $\mathsf{ssap} : \mathsf{Fmla}_1 \to \mathcal{F}_1$ correctly captures substituted self-application: $\mathsf{ssap}\,\psi\,\langle\textcircled{f}\rangle = \langle\psi(\textcircled{f}\langle\textcircled{f}\rangle)\rangle$ for all $\psi \in \mathsf{Fmla}_1$ and $f \in \mathcal{F}_1$.

In CapSS, following Jeroslow we employed Notation 14 taking advantage of the fact that $\textcircled{f}$ are pseudo-terms: The highlighted text denotes $\exists y.\ \textcircled{f}(\langle\textcircled{f}\rangle, y) \wedge \psi(y)$. Moreover, using the same notation, the statement of $\mathsf{Repr}_{\mathcal{F}}$ for some $f \in \mathcal{F}_1$ and $n \in \mathsf{Num}$ would be written as $\vdash f(n) \equiv \textcircled{f}(n)$. Similarly, combining CapN with the instance of $\mathsf{Repr}_{\mathcal{F}}$, we obtain a fact that can be written as $\vdash \langle\neg\,\varphi\rangle \equiv \textcircled{\mathsf{N}}\langle\varphi\rangle$.

When our logical theory is a recursive extension of the Robinson arithmetic and $\mathsf{Num} = \mathbb{N}$, $\mathcal{F}_m$ could be the set of m-ary computable functions. Then every $f \in \mathcal{F}_m$ would indeed be represented by a formula $\textcircled{f}$. Moreover, assuming a computable and injective encoding of formulas, $\langle\_\rangle : \mathsf{Fmla}_1 \to \mathbb{N}$, we can take $\mathsf{N} : \mathbb{N} \to \mathbb{N}$ to be the following computable function: Given input n, it checks if n has the form $\langle\varphi\rangle$; if so, it returns $\langle\neg\,\varphi\rangle$; if not, it returns any value (e.g., 0). And $\mathsf{ssap}\,\psi$ can be defined similarly, obtaining the desired property for every $\varphi \in \mathsf{Fmla}_2$, not necessarily of the form $\textcircled{f}$. In short, Jeroslow's assumptions cover arithmetic (but also potentially many other systems).

At the heart of Jeroslow's approach lies an alternative diagonalization technique, producing *term* fixpoints, not just formula fixpoints:

Lemma 15. Assume CapSS and $\mathsf{Repr}_{\mathcal{F}}$ and let $\psi \in \mathsf{Fmla}_1$. Then there exists a closed pseudo-term t such that $\vdash t \equiv \langle\psi(t)\rangle$. Moreover, taking $\varphi = \psi(t)$, we have $\vdash \varphi \leftrightarrow \psi\langle\varphi\rangle$.

Proof. Let $f = \mathsf{ssap}\,\psi$ and $t = \textcircled{f}\langle\textcircled{f}\rangle$. From CapSS, we obtain $f\langle\textcircled{f}\rangle = \langle\psi(\textcircled{f}\langle\textcircled{f}\rangle)\rangle$. From this and $\mathsf{Repr}_{\mathcal{F}}$, we obtain $\vdash \textcircled{f}\langle\textcircled{f}\rangle \equiv \langle\psi(\textcircled{f}\langle\textcircled{f}\rangle)\rangle$, i.e., $\vdash t \equiv \langle\psi(t)\rangle$. With the equality rules, we obtain $\vdash \psi(t) \leftrightarrow \psi(\langle\psi(t)\rangle)$, i.e., $\vdash \varphi \leftrightarrow \psi\langle\varphi\rangle$. □

This lemma offers us Gödel and Rosser sentences, which can be used like in Sections 3.1 and 3.2, leading to corresponding variants of $\mathcal{IT}_1$. But Jeroslow's main innovation affects $\mathcal{IT}_2$: While traditionally $\mathcal{IT}_2$ requires all three derivability conditions, Jeroslow's version does not make use of the second, HBL_2:

Theorem 16. ($\mathcal{IT}_2$ *à la* Jeroslow) Assume Con, HBL_1, SHBL_3, $\mathsf{Repr}_{\mathcal{F}}$, CapN, CapSS. Then $\nvdash \mathsf{jcon}$, where jcon denotes $\forall x.\ \neg\,(\textcircled{\vdash}(x) \wedge \textcircled{\vdash}(\textcircled{\mathsf{N}}(x)))$.

Like with Rosser's trick, we analyze this innovation's trade-offs from an abstract perspective. A first trade-off is in the employment of a stronger version of the third condition, SHBL_3 (extended to affect all closed pseudo-terms via Notation 14). Another is in the way consistency is expressed in the logic. Jeroslow does not conclude $\nvdash \neg\,\textcircled{\vdash}\langle\bot\rangle$, but something more elaborate, namely $\nvdash \mathsf{jcon}$. While the formula $\neg\,\textcircled{\vdash}\langle\bot\rangle$ internalizes the statement $\nvdash \bot$, jcon internalizes the equivalent statement "for all φ, it is not the case that $\vdash \varphi$ and $\vdash \neg\,\varphi$." But are the internalizations themselves equivalent, i.e., is it the case that $\vdash \neg\,\textcircled{\vdash}\langle\bot\rangle$ iff $\vdash \mathsf{jcon}$? This surely holds for many concrete logics, but it is one direction that we can infer logic-independently: Assuming HBL_1, $\mathsf{Repr}_{\mathcal{F}}$ and CapN, $\vdash \mathsf{jcon}$ implies $\vdash \neg\,\textcircled{\vdash}\langle\bot\rangle$. And it seems we cannot infer the other direction without knowing how $\textcircled{\vdash}$ looks like more concretely. Therefore, $\nvdash \neg\,\textcircled{\vdash}\langle\bot\rangle$, the conclusion of the original $\mathcal{IT}_2$, is *abstractly stronger than*, hence *preferable to* $\nvdash \mathsf{jcon}$. In short, Jeroslow somewhat weakens the theorem's conclusion.

Let us now look at (a slight rephrasing of) Jeroslow's proof:

Proof of Theorem 16. We assume (*) $\vdash \mathsf{jcon}$ and aim to reach a contradiction.
- Applying Lemma 15 to $\textcircled{\vdash}(\textcircled{\mathtt{N}}(x))$, we obtain the closed term t such that
(**) $\vdash t \equiv \langle\, \textcircled{\vdash}(\textcircled{\mathtt{N}}(t)) \,\rangle$.
- By SHBL_3 applied to $\textcircled{\mathtt{N}}(t)$, we obtain $\vdash \textcircled{\vdash}(\textcircled{\mathtt{N}}(t)) \to \textcircled{\vdash}\langle\textcircled{\vdash}(\textcircled{\mathtt{N}}(t))\rangle$.
- From (**) and the equality rules, we obtain $\vdash \textcircled{\vdash}(\textcircled{\mathtt{N}}(t)) \to \textcircled{\vdash}(\textcircled{\mathtt{N}}\langle\textcircled{\vdash}(\textcircled{\mathtt{N}}(t))\rangle)$.
- The last two facts give us $\vdash \varphi \to \textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}(\textcircled{\mathtt{N}}\langle\varphi\rangle)$, where φ denotes $\textcircled{\vdash}(\textcircled{\mathtt{N}}(t))$.
- On the other hand, (*) instantiated with $\langle\varphi\rangle$ gives us $\vdash \neg\,(\textcircled{\vdash}\langle\varphi\rangle \wedge \textcircled{\vdash}(\textcircled{\mathtt{N}}\langle\varphi\rangle))$.
- From the last two facts, we obtain (***) $\vdash \neg\,\varphi$.
- With HBL_1, we obtain $\vdash \textcircled{\vdash}\langle\neg\,\varphi\rangle$.
- With CapN and $\mathsf{Repr}_{\mathcal{F}}$, we obtain $\vdash \textcircled{\vdash}(\textcircled{\mathtt{N}}\langle\varphi\rangle)$.
- From (**) and the equality rules, we obtain $\vdash \textcircled{\vdash}(\textcircled{\mathtt{N}}\langle\textcircled{\vdash}(\textcircled{\mathtt{N}}(t))\rangle) \to \textcircled{\vdash}(\textcircled{\mathtt{N}}(t))$, i.e., $\vdash \textcircled{\vdash}(\textcircled{\mathtt{N}}\langle\varphi\rangle) \to \varphi$
- From the last two facts, we obtain $\vdash \varphi$.
- This and (***) contradict the consistency assumption. □

A first major observation is that, under the stated assumptions, the above proof is *incorrect*. It uses an implicit assumption, hidden under Notation 14: When we disam-

biguate the notation, we see that Lemma 15 gives us a pseudo-term t that does not exactly satisfy (1) $\vdash t \equiv \langle\psi(t)\rangle$ (which is what the theorem's proof needs), but something weaker, namely (2) $\vdash t \equiv \langle\chi\rangle$, where χ is $\vdash \exists x.\, ⓕ(\langle ⓕ\rangle, x) \wedge \psi(x)$. And although $\vdash \chi \leftrightarrow \psi(t)$, we still cannot infer (1) from (2), unless *the encodings of provably equivalent formulas are assumed provably equal*. But this assumption is unreasonable: Usually formula equivalence is undecidable, so no computable encoding can achieve that. (Incidentally, this problem is also the reason why we need SHBL_3 instead of HBL_3: In the proof's application of SHBL_3 to obtain $\vdash ⊢(ℕ(t)) \rightarrow ⊢\langle ⊢(ℕ(t))\rangle$, we cannot work with $\langle\neg\,\varphi\rangle$ instead of $ℕ(t)$, even though $\vdash \langle\neg\,\varphi\rangle \equiv ℕ(t)$.)

To repair that, we can replace representation by pseudo-terms with actual term-representation. Namely, if we change $\mathsf{Repr}_{\mathcal{F}}$ into "Every $f \in \mathcal{F}_m$ is term-represented by some $ⓕ : \mathsf{Term}^m \rightarrow \mathsf{Term}$ under the identity encoding $\mathsf{Num} \rightarrow \mathsf{Num}$," then CapSS, Lemma 15, and all proofs can operate with terms rather than pseudo-terms and everything will be formally correct. In summary, it seems that Jeroslow's approach to $\mathcal{IT}_2$ fails for pseudo-terms representing computable functions, but requires actual terms. This usually means that the logic must have built-in Skolem symbols and axioms.

Finally, let us see what it takes to alleviate the second trade-off: from $\nvdash \mathsf{jcon}$ to the more desirable $\nvdash \neg\, ⊢\langle\bot\rangle$. We see that Theorem 16's proof uses $\vdash \mathsf{jcon}$ not at jcon's full generality but only instantiated with formula encodings, which thanks to $\mathsf{Repr}_{\mathcal{F}}$ and CapN would follow from (*) $\vdash \neg\,(⊢\langle\varphi\rangle \wedge ⊢\langle\neg\,\varphi\rangle)$. And it only takes WHBL_2 (a weaker version of HBL_2) and HBL_4 to prove $\vdash (⊢\langle\varphi\rangle \wedge ⊢\langle\neg\,\varphi\rangle) \rightarrow ⊢\langle\bot\rangle$, allowing us to infer (*) from $\vdash \neg\, ⊢\langle\bot\rangle$; meaning that the latter could have been used. We obtain:

Theorem 17. If in the (corrected) Theorem 16 we additionally assume WHBL_2 and HBL_4, its conclusion can be upgraded to $\nvdash \neg\, ⊢\langle\bot\rangle$.

Whether WHBL_2 and HBL_4 are a good trade-off for HBL_2 will of course depend on the logic's specificity, in particular, on its primitive rules of inference.

Jeroslow presented his approach for an abstract logical theory over a FOL language, which is not necessarily a FOL theory—so it found a natural fit in our generic framework. To our knowledge, very few subsequent authors present Jeroslow's approach rigorously, and none at its original level of generality. Smith's monograph gives a rigorous account for arithmetic [32, §33], silently performing the correction we have shown here, but failing to detect the need for SHBL_3 instead of HBL_3 (which Jeroslow did not). A mechanical prover is of invaluable help for detecting such nuances and pitfalls.

We conclude with an anecdote involving our prover and Jeroslow's notations. Given the simplicity of Lemma 15 (version with terms, not pseudo-terms), it came as no surprise that Isabelle's Sledgehammer [26] was able to prove it automatically. Sledgehammer reported to have used $\vdash$-reflexivity in the proof. And indeed, it had found a term t for which it had proved not just $\vdash t \equiv \langle\psi(t)\rangle$, but actual equality, $t = \langle\psi(t)\rangle$; in particular, the term was a numeral. All this was too good to be true. It took us some time to realize why that happened: Due to another one of Jeroslow's notations, who wrote f instead of $ⓕ$ (thus identifying a function with its representing pseudo-term), we had at first misstated CapSS, writing $\langle\psi(f\langle ⓕ\rangle)\rangle$ instead of $\langle\psi(ⓕ\langle ⓕ\rangle)\rangle$; the former is still a valid expression, since f is a function between numerals which are particular terms. Embarrassingly, it took us even longer to realize why this variation discovered by chance was not an improvement of Jeroslow's diagonalization lemma: because the

assumption CapSS becomes unreasonable. Indeed, no concrete computable function can act like the intended $\mathsf{ssap}\,\psi$: Given an input n, (1) decode it into a unique formula φ such that $n = \langle\varphi\rangle$, (2) decode φ into a unique function f such that $\varphi = \textcircled{f}$ and (3) proceed to apply f as part of producing $\langle\psi(f\langle\textcircled{f}\rangle)\rangle$. The second step requires an injective and computable encoding of computable functions into formulas, which is impossible.

Summary Using our generic infrastructure (Section 2), we have formally proved several abstract incompleteness results. They include four versions of $\mathcal{IT}_1$:

- Gödel's original $\mathcal{IT}_1$ (Theorem 9) and an $\mathcal{IT}_1$ based on classical logic (Theorem 12) required the formalization of some well-known arguments without change.
- Rosser's $\mathcal{IT}_1$ (Theorem 10) involved the generalization of a well-known argument: distilling two abstract conditions, Ord_1 and Ord_2.
- A novel semantic $\mathcal{IT}_1$ (Theorem 11) was born from analyzing the relationship between standard models, the "iff" version of HBL_1, and proof representability.

They also include two versions of $\mathcal{IT}_2$:

- The standard $\mathcal{IT}_2$ based on the three derivability conditions (Theorem 13) again only required formalizing a well-known argument.
- The alternative, Jeroslow-style $\mathcal{IT}_2$ (Theorems 16 and 17) involved a detailed analysis and correction of an existing abstract result.

4 Instances of the Abstract Results

We first validate the assumptions about our abstract logic and arithmetic:

Prop 18. (1) Any FOL theory that extends the Robinson arithmetic or the HF set theory satisfies all the axioms in our logical and arithmetical substrata (in Sections 2.1 and 2.2). (2) If, in addition, the theory is sound, then, together with its corresponding standard model, it also satisfies all our model-theoretic axioms (in Section 2.5).

In particular, point (2) shows that our discussion of standard models applies equally well to $\mathbb{N}$ and the datatype of HF sets. (In the latter case, Num becomes the entire set of closed terms, so that numerals can denote arbitrary HF sets. This shows the versatility of our abstract concept of numeral.) Then we instantiate three of our main theorems:

Theorem 19. (1) Any FOL theory that extends the HF set theory with a finite set of axioms and is sound in the standard HF set model satisfies the hypotheses of Theorems 11 and 13. Hence $\mathcal{IT}_1$ (semantic version) and $\mathcal{IT}_2$ hold for it.
(2) Any FOL theory that extends the HF set theory with a finite set of axioms and is consistent satisfies the hypotheses of Theorem 13. Hence $\mathcal{IT}_2$ holds for it.

These instances are heavily based on the lemmas proved by Paulson in his formalization of $\mathcal{IT}_1$ and $\mathcal{IT}_2$ [24, 25], who follows and corrects Świerczkowski's detailed informal account [34]. Point (1) is a restatement of Paulson's formalized results: theorems *Goedel_I* and *Goedel_II* in [25]. (His theorems also assume consistency, but that is redundant: Consistency follows from his underlying soundness assumption.)

By contrast, point (2) is an upgrade of Paulson's *Goedel_II*, applicable to any consistent, though possibly unsound theory. This stronger version is in fact $\mathcal{IT}_2$'s standard

form, free from any model-theoretic considerations.[3] Paulson had proved both HBL_1 and $\mathsf{HBL}_1^{\Leftarrow}$ taking advantage of soundness, so we needed to discard $\mathsf{HBL}_1^{\Leftarrow}$ and re-prove HBL_1 by replacing any semantic arguments with proofs within the HF calculus. We also removed all invocations of a convenient "truth implies provability for Σ-sentences" lemma, which depended on soundness due to Paulson's choice of Σ-sentence definition.

This instantiation process has offered important feedback into the abstract results. A formal development such as ours is (largely) immune to reasoning errors, but not to missing out on useful pieces of generality. We experienced this firsthand with our assumptions about substitution. An *a priori* natural choice was to assume representability of the numeral substitution $\mathsf{Sb} : \mathsf{Fmla}_1 \times \mathsf{Num} \to \mathsf{Sen}$ (defined as $\mathsf{Sb}(\varphi, n) = \varphi(n)$), part of which means (1) $\vdash \mathfrak{Sb}(\langle\varphi\rangle, n, \mathsf{Sb}(\varphi, n))$. But Paulson had instead proved (2) $\vdash \mathfrak{Sb}(\langle\varphi\rangle, \langle n\rangle, \mathsf{Sb}(\varphi, n))$. The key difference from (1) is that (2) applies the term encoding function $\langle\_\rangle : \mathsf{Term} \to \mathsf{Num}$ to numerals as well (as particular terms); and since his $\langle\_\rangle$ function is injective, it is far from the case that $\langle n\rangle = n$ for all numerals n. Paulson's version makes more sense than ours when building the results bottom-up: Representability should not discriminate numerals, but filter them through the encodings like other terms. However, top-down our version also made sense: It yielded the incompleteness theorems under reasonable assumptions, which do hold, by the way, for the HF set theory—even though in a bottom-up development one is unlikely to prove them. We resolved this discrepancy through a common denominator: the representability of self-substitution $\mathsf{S} : \mathsf{Fmla}_1 \to \mathsf{Sen}$ (Section 2.3), which made our results more general.

Paulson's formalization has also inspired our abstract treatment of standard models (Section 2.5). Since Paulson proves $\mathsf{HBL}_1^{\Leftarrow}$ and uses classical logic, an obvious "port of entry" of his $\mathcal{IT}_2$ into our framework is Theorem 12. But this theorem tells us nothing about the truth of the Gödel sentences. Delving deeper into Paulson's proof, we noted that he (unconventionally) completely avoids $\mathsf{Repr}_{\Vdash}$, and does not even define $\Vdash$. This raised the question of whether $\mathsf{HBL}_1^{\Leftarrow}$ and $\mathsf{Repr}_{\Vdash}$ are somehow interchangeable in the presence of standard models—and we found that they indeed are, under mild assumptions about truth. Incidentally, these assumptions were also sufficient for establishing the Gödel sentences' truth, leading to our semantic $\mathcal{IT}_1$ (Theorem 11), which we believe expresses the essence of the Świerczkowski–Paulson approach.

Many other logics and logical theories satisfy our theorems' assumptions. We do *not* require the logic to be reducible to a single syntactic category of formulas, Fmla, a single syntactic judgment, $\vdash$, etc.; but only that such (well-behaved) formulas, provability relation, etc. are identifiable as part of that logic, e.g., localized to a given type and/or relativised by a given predicate. This allows our framework to capture most variants of higher-order logic and type theory, and also, we believe, many of the logics surveyed by Buldt [4], including non-classical and fuzzy. But enabling "mass instantiation" that is both formal and painless requires more progress on the agenda we started here: recognizing reusable construction and proof patterns and formalizing them as abstract results.

Acknowledgments. We are grateful to Bernd Buldt for his patient explanations on some of the material in his monograph.

[3] The value of freeing $\mathcal{IT}_2$ from its reliance on models becomes even greater (1) for expressive systems such as ZF set theory and (2) for systems that lack obvious notions of (standard) models—such as the logic of Isabelle/HOL itself [19, 20].

References

1. Auerbach, D.: Intensionality and the Gödel theorems. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 48(3), 337–351 (1985)
2. Blanchette, J.C., Popescu, A., Traytel, D.: Unified classical logic completeness—A coinductive pearl. In: Demri, S., Kapur, D., Weidenbach, C. (eds.) IJCAR 2014. LNCS, vol. 8562, pp. 46–60. Springer (2014)
3. Boolos, G.: The Logic of Provability. Cambridge University Press (1993)
4. Buldt, B.: The scope of Gödel's first incompleteness theorem. Logica Universalis 8(3), 499–552 (2014)
5. Carnap, R.: Logische syntax der sprache. Philosophical Review 44(4), 394–397 (1935)
6. Davis, M.: The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems, and Computable Functions. Dover Publication (1965)
7. Diaconescu, R.: Institution-independent Model Theory. Birkhäuser, 1st edn. (2008)
8. Feferman, S., Dawson, Jr., J.W., Kleene, S.C., Moore, G.H., Solovay, R.M., van Heijenoort, J. (eds.): Kurt Gödel: Collected Works. Vol. 1: Publications 1929-1936. Oxford University Press (1986)
9. Fiore, M.P., Plotkin, G.D., Turi, D.: Abstract syntax and variable binding. In: Logic in Computer Science (LICS) 1999, pp. 193–202. IEEE Computer Society (1999)
10. Gabbay, M.J., Mathijssen, A.: Nominal (universal) algebra: Equational logic with names and binding. J. Log. Comput. 19(6), 1455–1508 (2009)
11. Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik 38(1), 173–198 (1931)
12. Goguen, J.A., Burstall, R.M.: Institutions: Abstract model theory for specification and programming. J. ACM 39(1), 95–146 (1992)
13. Harrison, J.: HOL Light proof of Gödel's first incompleteness theorem, located at `http://code.google.com/p/hol-light/`, directory Arithmetic
14. Hilbert, D., Bernays, P.: Grundlagen der Mathematik, Vol. II (1939)
15. Jeroslow, R.G.: Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem. J. Symb. Log. 38(3), 359–367 (1973)
16. Kaliszyk, C., Urban, J.: HOL(y)Hammer: Online ATP service for HOL light. Mathematics in Computer Science 9(1), 5–22 (2015)
17. Kikuchi, M., Kurahashi, T.: Generalizations of Gödel's incompleteness theorems for $\sum$ n -definable theories of arithmetic. Rew. Symb. Logic 10(4), 603–616 (2017)
18. Kreisel, G.: Mathematical logic. In: Saaty, T.L. (ed.) Lectures on modern mathematics, vol. 3. Wiley (1963)
19. Kunčar, O., Popescu, A.: A Consistent Foundation for Isabelle/HOL. In: ITP. pp. 234–252 (2015)
20. Kunčar, O., Popescu, A.: Comprehending Isabelle/HOL's consistency. In: ESOP. pp. 724–749 (2017)
21. Löb, M.: Solution of a Problem of Leon Henkin. The Journal of Symbolic Logic 20(2), 115—-118 (1955)
22. Nipkow, T., Paulson, L., Wenzel, M.: Isabelle/HOL — A Proof Assistant for Higher-Order Logic, LNCS, vol. 2283. Springer (2002)
23. O'Connor, R.: Essential incompleteness of arithmetic verified by Coq. In: TPHOLs. pp. 245–260 (2005)
24. Paulson, L.C.: A machine-assisted proof of Gödel's incompleteness theorems for the theory of hereditarily finite sets. Rew. Symb. Logic 7(3), 484–498 (2014)
25. Paulson, L.C.: A mechanised proof of Gödel's incompleteness theorems using Nominal Isabelle. J. Autom. Reasoning 55(1), 1–37 (2015)

26. Paulson, L.C., Blanchette, J.C.: Three years of experience with Sledgehammer, a practical link between automatic and interactive theorem provers. In: The 8th International Workshop on the Implementation of Logics, IWIL 2010, Yogyakarta, Indonesia, October 9, 2011. pp. 1–11 (2010)
27. Popescu, A., Rosu, G.: Term-generic logic. Theor. Comput. Sci. 577, 1–24 (2015)
28. Popescu, A., Traytel, D.: Formalization associated with this paper. `https://bitbucket.org/traytel/abstract_incompleteness/` (2019)
29. Raatikainen, P.: Gödel's incompleteness theorems. In: The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University (2018)
30. Schlichtkrull, A., Blanchette, J.C., Traytel, D., Waldmann, U.: Formalizing Bachmair and Ganzinger's ordered resolution prover. In: Galmiche, D., Schulz, S., Sebastiani, R. (eds.) IJCAR 2018. LNCS, vol. 10900, pp. 89–107. Springer (2018)
31. Shankar, N.: Metamathematics, Machines, and Gödel's Proof. Cambridge University Press (1994)
32. Smith, P.: An introduction to Gödel's incompleteness theorems. Cambridge University Press (2007)
33. Smorynski, C.: The incompleteness theorems. In: Barwise, J. (ed.) Handbook of Mathematical Logic, pp. 821–865. North-Holland (1977)
34. Świerczkowski, S.: Finite sets and Gödel's incompleteness theorems. Dissertationes Mathematicae 422, 1–58 (2003)
35. Traski, A., Mostowski, A., Robinson, R.: Undecidable Theories. Studies in Logic and the Foundations of Mathematics. North-Holland (1953), 3rd edition, 1971
36. Troelstra, A.S., Schwichtenberg, H.: Basic Proof Theory. Cambridge University Press (1996)

APPENDIX

This appendix contains more details on the concepts discussed in the main paper (Section A) and proof sketches for all the abstract results (Section B). The proof sketches have been provided for readers who are interested in the main line of reasoning but do not wish to inspect the formal Isabelle scripts.

A More details about various concepts

A.1 Free variables and substitution

The free-variable and substitution operators on terms and formulas have the following types, where $\mathcal{P}_{\mathsf{fin}}(\mathsf{Var})$ denotes the set of finite subsets of Var.

- $\mathsf{FVars} : \mathsf{Term} \to \mathcal{P}_{\mathsf{fin}}(\mathsf{Var})$
- $\mathsf{FVars} : \mathsf{Fmla} \to \mathcal{P}_{\mathsf{fin}}(\mathsf{Var})$
- $\_[\_/\_] : \mathsf{Term} \times \mathsf{Term} \times \mathsf{Var} \to \mathsf{Term}$
- $\_[\_/\_] : \mathsf{Fmla} \times \mathsf{Term} \times \mathsf{Var} \to \mathsf{Fmla}$

We do not assume a particular syntax for terms or formulas. All we assume is that the above operators on the abstarct sets Term and Fmla satisfy some properties:

FVV: Free-variables and substitution on variables behave as expected:
- $\mathsf{FVars}(x) = \{x\}$,
- $x\,[s/x] = s$, and $y\,[s/x] = y$ if $x \neq y$

SVac: Substitution on terms is vacuous outside the free variables: $x \notin \mathsf{FVars}(t)$ implies $t\,[s/x] = t$, and similarly for substitution on formulas

In addition, for the formula operators we assume the following:

FVS: Free-variables distribute over substitution:
$\mathsf{FVars}\,(\varphi\,[s/x]) = \mathsf{FVars}(\varphi) - \{x\} \cup \mathsf{FVars}(s)$ if $x \in \mathsf{FVars}(\varphi)$

SVacV: Substitution of a variable for itself is vacuous: $\varphi\,[x/x] = \varphi$

SC: Substitution composes with itself under certain freshness assumptions:
- $\varphi\,[s_1/x]\,[s_2/x] = \varphi\,[(s_1[s_2/x])\ /\ x]$
- $\varphi\,[s_1/x_1]\,[s_2/x_2] = \varphi\,[(s_1[s_2/x_2])\ /\ x_1]$ if $x_2 \notin \mathsf{FVars}(\varphi)$
- $\varphi\,[s_1/x_1]\,[s_2/x_2] = \varphi\,[s_2/x_2]\,[(s_1[s_2/x_2])\ /\ x_1]$ if $x_1 \neq x_2$ and $x_1 \notin \mathsf{FVars}(s_2)$

Of the above assumptions, FVV only applies to, and only makes sense for, substitution on terms. By contrast, we assume SVac for both terms and formulas. The last group, FVS, SVacV and SC, would makes sense for terms too, but we only assume them for formulas—this is in line with our Economy principle, since we did not need them for terms.

A.2 Numerals

Numerals are assumed to have no free variables:

FVNum: $\mathsf{FVars}(n) = \emptyset$ for all $n \in \mathsf{Num}$.

By SVac, this implies that substitution on numerals is vacuous.

A.3 Abstract connectives and quantifiers

The main theorems depend on the existence of (subsets of) the following connectives and quantifiers, which we represent as operators on terms:

- Equality, $\equiv : \mathsf{Term} \to \mathsf{Term} \to \mathsf{Fmla}$
- False, $\bot \in \mathsf{Fmla}$
- Conjunction, implication and disjunction, $\vee, \wedge, \to : \mathsf{Fmla} \times \mathsf{Fmla} \to \mathsf{Fmla}$
- Universal and existential quantification, $\forall, \exists : \mathsf{Var} \times \mathsf{Fmla} \to \mathsf{Fmla}$

When we need negation, $\neg$, we define it from implication as False, taking $\neg\,\varphi$ to be $\varphi \to \bot$. On the other hand, even in the presence of negation, we do not assume that $\vee$ and $\exists$ are definable from $\wedge$ and $\forall$ or vice versa—this is because, in line with the Economy principle, we use intuitionistic, not classical logic.

A.4 Provability

In our formalization, we prefer to start with a Hilbert-style system because this is the most economic (requires the least amount of infrastructure). Namely, we consider the assumptions corresponding to the standard implication-based Hilbert system for intuitionistic logic [36]:

$\vdash \varphi \to \psi$ and $\vdash \varphi$ implies $\vdash \psi$
$\vdash \varphi \to (\psi \to \varphi)$
$\vdash (\varphi \to (\psi \to \chi)) \to (\varphi \to \psi) \to (\varphi \to \chi)$

Then each connective is integrated into this system on a need basis—if it is used in the theorems. For example, here are the assumptions for conjunction

$\vdash \varphi \wedge \psi \to \varphi$
$\vdash \varphi \wedge \psi \to \psi$
$\vdash \varphi \to (\psi \to \varphi \wedge \psi)$

and here are those for universal quantification:

$x \notin \mathsf{FVars}(\varphi)$ and $\vdash \varphi \to \psi$ implies $\vdash \varphi \to (\forall x.\, \psi)$
$\vdash (\forall x.\, \varphi) \to \varphi[t/x]$

And similarly for existential, equality, etc. Notice again that we do take the above to be an inductive definition of $\vdash$, but simply assume that $\vdash$ is closed under those rules. In particular, the system for classical logic satisfies our assumptions.

Since a natural deduction system is more convenient for performing formal deductions, in our formalization we extract one from $\vdash$. Namely, in the presence of conjunction and implication, we extend $\vdash$ to a binary relation $\_ \vdash \_$ between finite sets F and formulas and single formulas φ, defined as $F \vdash \varphi$ iff $\vdash (\bigwedge F) \to \varphi$. We think of $F \vdash \varphi$ as "φ follows from F." For this relation, we prove the standard introduction and elimination rules for each connective and quantifier.

A.5 ω-consistency

Lemma 20. Assume classical deduction. Then OCon is equivalent to: For all $\varphi \in \mathsf{Fmla}_1^{\subseteq}$, it is not the case that $\vdash \varphi(n)$ for all $n \in \mathsf{Num}$ and $\vdash \neg\, (\forall x.\, \varphi(x))$.

Proof. Immediate, switching between φ to $\neg\, \varphi$ and using the double negation axiom. □

A.6 Order-like relation

In the presence of a numeral-restricted form of anti-symmetry of the relation (which would include $<$ but exclude $\leq$), the second condition is stronger:

Lemma 21. Assume $\vdash \forall x.\, x \prec n \to \neg\, n \prec x$. Then Ord_2 implies Ord_1.

Proof. Let $\varphi \in \mathsf{Fmla}_1$ and $n \in \mathsf{Num}$. Assume $\vdash \varphi(m)$ for all $m \in \mathsf{Num}$ and let M be as in Ord_2. We must prove $\vdash \forall x \prec n.\, \varphi(x)$. Working inside the formal proof system $\vdash$, we fix x such that $x \prec n$. Thanks to the antisymmetry assumption, we obtain $\neg\, n \prec x$, which implies by Ord_2 that x equals some $m \in M$; this means that $\varphi(x)$ holds, as desired. □

A.7 Clean representations

The key property of cleanness is that it makes a representation behave well with respect to universal quantification of negative statements. We illustrate this for the binary case:

Lemma 22. Assume $R \subseteq A \times B$ is represented by $\textcircled{R}$, and this representation is 1-clean. Then the following are equivalent:

(1) $(a, b) \notin R$ for all $a \in A$ (2) $\vdash \neg\, \textcircled{R}(n, \langle b \rangle)$ for all $n \in \mathsf{Num}$

Proof. By representability, (1) is equivalent to: (i) $\vdash \neg\, \textcircled{R}(\langle a \rangle, \langle b \rangle)$ for all $a \in A$. Thanks to 1-cleanness, we can exclude the numerals outside the image of $\langle \_ \rangle$, obtaining that (i) is equivalent to (2). □

A.8 Weaker version of the second derivability condition

Lemma 23. HBL_1 and HBL_2 imply WHBL_2

Proof. Assume $\vdash \varphi \to \psi$.
- With[4] HBL_1, we obtain $\vdash \textcircled{\vdash}\langle \varphi \to \psi \rangle$.
- From HBL_2, we have $\vdash \textcircled{\vdash}\langle \varphi \to \psi \rangle \to (\textcircled{\vdash}\langle \varphi \rangle \to \textcircled{\vdash}\langle \psi \rangle)$.
- From the last two facts, we have $\vdash \textcircled{\vdash}\langle \varphi \rangle \to \textcircled{\vdash}\langle \psi \rangle$, as desired. □

A.9 HBL_1 from representability of "proof of"

Remember that we take $\textcircled{\vdash} \in \mathsf{Fmla}_1$ to be $\exists x.\, \textcircled{\Vdash}(x, v_1)$.

Lemma 24. $\mathsf{Rel}_{\vdash}^{\Vdash}$ and $\mathsf{Repr}_{\Vdash}$ imply HBL_1.

Proof. Assume $\vdash \varphi$. Then there exists $p \in \mathsf{Proof}$ such that $p \Vdash \varphi$. By $\mathsf{Repr}_{\Vdash}$, we have $\vdash \textcircled{\Vdash}(\langle p \rangle, \langle \varphi \rangle)$, hence $\exists x.\, \textcircled{\Vdash}(x, \langle \varphi \rangle)$, as desired. (Note that we did not need the whole $\mathsf{Repr}_{\Vdash}$; the weak representability of $\Vdash$ would have sufficed.) □

[4] Here and in all the other proofs, by "With …" we mean "From this (fact) and …".

A.10 More on Rosser's variation of Gödel's First

The following propositions cover the two halves of Rosser's $\mathcal{IT}_1$ (Theorem 10). The next one works with Rosser sentences instead of Gödel sentences, in order to avoid Prop. 8's ω-consistency assumption:

Prop 25. Assume Con, Ord_2, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$ $\mathsf{Repr}_{\neg}$ and $\mathsf{Clean}_{\Vdash}$. Then $\nvdash \neg \mathsf{R}$ for all Rosser sentences R.

Proof. To prove $\nvdash \neg \mathsf{R}$, we assume (1) $\vdash \neg \mathsf{R}$ and aim to reach a contradiction.
- By $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $p \Vdash \neg \mathsf{R}$ for some $p \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$, we obtain (2) $\vdash \textcircled{\scriptsize P}(\langle p \rangle, \langle \neg \mathsf{R} \rangle)$.
- From (1) and consistency, we obtain $\nvdash \mathsf{R}$.
- With $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $q \not\Vdash \mathsf{R}$ for all $q \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ and Lemma 22, we obtain (3) $\vdash \neg \textcircled{\scriptsize P}(n, \langle \mathsf{R} \rangle)$ for all $n \in \mathsf{Num}$.
- By Ord_2, we obtain a finite $M \subseteq \mathsf{Num}$ such that (4) $\forall x.\ \vdash (\bigvee_{m \in M} x \equiv m) \vee \langle p \rangle \prec x$
- We prove (5) $\vdash \forall x.\ \neg\, (\textcircled{\scriptsize P}(x, \langle \mathsf{R} \rangle) \wedge \mathsf{RosserTwist}(x, \langle \mathsf{R} \rangle))$. The proof is performed in our intuitionistic proof system, but we describe it informally: Fixing x, we assume $\textcircled{\scriptsize P}(x, \langle \mathsf{R} \rangle) \wedge \mathsf{RosserTwist}(x, \langle \mathsf{R} \rangle)$ and must reach a contradiction. We perform a case distinction according to (4):
— If x equals some $m \in M$, then $\textcircled{\scriptsize P}(m, \langle \mathsf{R} \rangle)$, which together with (3) leads to contradiction.
— If $\langle p \rangle \prec x$, then from $\mathsf{RosserTwist}(x, \langle \mathsf{R} \rangle)$ and $\textcircled{\scriptsize N}(\langle \mathsf{R} \rangle, \langle \neg \mathsf{R} \rangle)$ (which holds thanks to $\mathsf{Repr}_{\neg}$), we obtain $\neg\, \textcircled{\scriptsize P}(\langle p \rangle, \langle \neg \mathsf{R} \rangle)$, which contradicts (2).
- From (5), by (intuitionistic) logic we obtain $\vdash \neg\, \exists x.\ (\textcircled{\scriptsize P}(x, \langle \mathsf{R} \rangle) \wedge \mathsf{RosserTwist}(x, \langle \mathsf{R} \rangle))$.
- Thanks to R being a Rosser formula, we obtain $\vdash \mathsf{R}$.
- Together with (1), this contradicts consistency. □

Thus, ω-consistency (assumption OCon) has been indeed weakened to consistency (assumption Con), but in exchange we needed to additionally assume a special formula $\prec$ satisfying Ord_2. This represents a quite strong commitment to the arithmetical ordering.

Even worse, this fix on the assumptions needed to show the unprovability of the negated formula ($\neg \mathsf{R}$) complicates the proof of the unprovability of the *direct* formula (R), which was trivial in Gödel's version (Prop. 7). Unlike in Gödel's version, we again need a cleanly representable "proof of" relation, representable negation, and well-behavedness of the order-like relation $\prec$:

Prop 26. Assume Con, Ord_1, $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$ $\mathsf{Repr}_{\neg}$ and $\mathsf{Clean}_{\Vdash}$. Then $\nvdash \mathsf{R}$ for all Rosser sentences R.

Proof. To prove $\nvdash \mathsf{R}$, we assume (1) $\vdash \mathsf{R}$ and aim to reach a contradiction.
- With $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $p \Vdash \mathsf{R}$ for some $p \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$, we obtain (2) $\vdash \textcircled{\scriptsize P}(\langle p \rangle, \langle \mathsf{R} \rangle)$.
- With (1) and consistency, we obtain $\nvdash \neg \mathsf{R}$.
- With $\mathsf{Rel}^{\Vdash}_{\vdash}$, we obtain $q \not\Vdash \neg \mathsf{R}$ for all $q \in \mathsf{Proof}$.
- With $\mathsf{Repr}_{\Vdash}$, $\mathsf{Clean}_{\Vdash}$ and Lemma 22, we have $\vdash \neg \textcircled{\scriptsize P}(n, \langle \neg \mathsf{R} \rangle)$ for all $n \in \mathsf{Num}$.
- With (2) and Ord_1, we obtain $\vdash \forall y \prec \langle p \rangle.\ \neg \textcircled{\scriptsize P}(y, \langle \neg \mathsf{R} \rangle)$.

- Since, by $\mathsf{Repr}_\neg$, the only z such that $\ominus(z, \langle\neg\,\mathsf{R}\rangle)$ is $\langle\neg\,\mathsf{R}\rangle$, we obtain $\vdash \mathsf{RosserTwist}(\langle p\rangle, \langle\mathsf{R}\rangle)$.
- From this and (2), we obtain $\vdash \exists x.\, \oplus(x, \langle\mathsf{R}\rangle) \wedge \mathsf{RosserTwist}(x, \langle\mathsf{R}\rangle)$.
- But since R is a Rosser sentence, from (1) we obtain $\vdash \neg\, \exists x.\, \oplus(x, \langle\mathsf{R}\rangle) \wedge \mathsf{RosserTwist}(x, \langle\mathsf{R}\rangle)$.
- The last two facts contradict consistency. □

A.11 Jeroslow's approach

Here is Jeroslow's version of Gödel's First:

Theorem 27. Theorems 9–12 still hold if we replace the Repr_S assumption with CapSS and $\mathsf{Repr}_\mathcal{F}$.

Proof. Lemma 15 allows to construct and use Gödel and Rosser sentences just like in Sections 3.1 and 3.2. □

The following is claimed in the main paper:

Lemma 28. Assume HBL_1, $\mathsf{Repr}_\mathcal{F}$, CapN. Then $\vdash \mathsf{jcon}$ implies $\vdash \neg\, ⊢\langle\bot\rangle$.

Proof. Assume $\vdash \mathsf{jcon}$.
- From this, we obtain (1) $\vdash \mathsf{N}(\langle\bot\rangle, \langle\neg\,\bot\rangle) \to \neg\, (⊢\langle\bot\rangle \wedge ⊢\langle\neg\,\bot\rangle)$.
- From $\mathsf{Repr}_\mathcal{F}$ and CapN, we obtain $\vdash \mathsf{N}(\langle\bot\rangle, \langle\neg\,\bot\rangle)$.
- With (1), we obtain (2) $\vdash \neg\, (⊢\langle\bot\rangle \wedge ⊢\langle\neg\,\bot\rangle)$.
- From $\vdash \neg\,\bot$ and HBL_1, we obtain $\vdash ⊢\langle\neg\,\bot\rangle$.
- With (2), we obtain $\vdash \neg\, (⊢\langle\bot\rangle)$, as desired. □

B Other proof sketches

Lemma 3

Proof. (1) Assume $\vdash ⊢\langle\varphi\rangle$.
- Hence $\vdash \exists x.\, ⊩(x, \langle\varphi\rangle)$.
- By (intuitionistic) logic, we obtain $\vdash \neg\,\neg\, (\exists x.\, ⊩(x, \langle\varphi\rangle))$.
- With OCon, we obtain $n \in \mathsf{Num}$ such that $\nvdash \neg\, ⊩(n, \langle\varphi\rangle)$.
- With $\mathsf{Clean}_{⊩}$, we obtain $p \in \mathsf{Proof}$ such that $n = \langle p\rangle$. Hence $\nvdash \neg\, ⊩(\langle p\rangle, \langle\varphi\rangle)$.
- Since, by $\mathsf{Repr}_{⊩}$, we have that $p \nVdash \varphi$ implies $\vdash \neg\, ⊩(\langle p\rangle, \langle\varphi\rangle)$, we obtain $p \Vdash \varphi$.
- With $\mathsf{Rel}^{⊩}_{⊢}$, we obtain $\vdash \varphi$, as desired.
(2): Assume $\vdash \neg\, \varphi(n)$ for all $n \in \mathsf{Num}$.
- With Soundness, we obtain $\models \neg\, \varphi(n)$ for all $n \in \mathsf{Num}$.
- With $\mathsf{Syn}_\models$(3), we obtain $\models \forall x.\, \neg\, \varphi(x)$.
- With the intuitionistic logic built in $\vdash$ and Soundness, we obtain $\models \neg\, (\exists x.\, \varphi(x))$.
- With $\mathsf{Syn}_\models$(1,2), we obtain $\not\models \neg\,\neg\, (\exists x.\, \varphi(x))$.
- With Soundness, we obtain $\nvdash \neg\,\neg\, (\exists x.\, \varphi(x))$, as desired.
(3): Assume $\models ⊢\langle\varphi\rangle$.
- Then $\models \exists x.\, ⊩(x, \langle\varphi\rangle)$.
- With $\mathsf{Syn}_\models$(4), we obtain $n \in \mathsf{Num}$ such that (i) $\models ⊩(n, \langle\varphi\rangle)$.

- With $\mathsf{Soundness}$ and $\mathsf{Syn}_{\models}$(1,2, we obtain $\nvdash \neg \textcircled{\textsf{P}}(n, \langle\varphi\rangle)$.
- From here, the proof of $\vdash \varphi$ proceeds just like at point (1): using $\mathsf{Rel}^{\Vdash}_{\vdash}$, $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$. □

Lemma 4

Proof. To show $\mathsf{Rel}^{\Vdash}_{\vdash}$ in this context (that is, for this particular definitions of Proof and relation $\Vdash$), we must show the equivalence between (i) $\vdash \varphi$ and (ii) the existence of $n \in \mathsf{Num}$ such that $\vdash \mathsf{P}\ (n, \langle\varphi\rangle)$.
First assume (i).
- With HBL_1, we obtain $\vdash \textcircled{\textsf{T}}\langle\varphi\rangle$.
- With $\mathsf{Rel}^{\mathsf{P}}_{\textcircled{\textsf{T}}}$, we obtain $\vdash \exists x.\ \mathsf{P}\ (x, \langle\varphi\rangle)$.
- With $\mathsf{Soundness}$, we obtain $\models \exists x.\ \mathsf{P}\ (x, \langle\varphi\rangle)$.
- With $\mathsf{Syn}_{\models}$(4), we obtain $n \in \mathsf{Num}$ such that $\models \mathsf{P}\ (n, \langle\varphi\rangle)$.
- With $\mathsf{Compl}_{\mathsf{P}}$, we obtain (ii), as desired.
Now assume (ii).
- By logic, we obtain $\vdash \exists x.\ \mathsf{P}\ (x, \langle\varphi\rangle)$.
- With $\mathsf{Rel}^{\mathsf{P}}_{\textcircled{\textsf{T}}}$, we obtain $\vdash \textcircled{\textsf{T}}\langle\varphi\rangle$.
- With $\mathsf{HBL}_1^{\Leftarrow}$, we obtain (i), as desired.

Showing half of $\mathsf{Repr}_{\Vdash}$ in this context is trivial, as it amounts to showing that "$\vdash \mathsf{P}\ (n, \langle\varphi\rangle)$ implies $\vdash \mathsf{P}\ (n, \langle\varphi\rangle)$". For the other half, assume $\nvdash \mathsf{P}\ (n, \langle\varphi\rangle)$.
- With $\mathsf{Compl}_{\mathsf{P}}$, we obtain $\not\models \mathsf{P}\ (n, \langle\varphi\rangle)$.
- With $\mathsf{Syn}_{\models}$(5), we obtain $\models \neg\ \mathsf{P}\ (n, \langle\varphi\rangle)$.
- With $\mathsf{Compl}_{\neg\mathsf{P}}$, we obtain $\vdash \neg\ \mathsf{P}\ (n, \langle\varphi\rangle)$, as desired.

Finally $\mathsf{Clean}_{\Vdash}$ is trivial in this context, since the encoding of proofs is the identity. □

Prop 5

Proof. Assume $\mathsf{Repr}_{\mathsf{S}}$, where S is the "hard" self-substitution function. Let $\chi \in \mathsf{Fmla}_1$ be $\exists y.\ \textcircled{\textsf{S}}(x, y) \wedge \psi(y)$. We take φ to be $\chi\langle\chi\rangle$.
- We must prove (1) $\vdash \varphi \leftrightarrow \psi\langle\varphi\rangle$.
- Using the definition of χ, this further means (2) $\vdash (\exists y.\ \textcircled{\textsf{S}}(\langle\chi\rangle, y) \wedge \psi(y)) \leftrightarrow \psi\langle\varphi\rangle$.
- By the fact that S is represented by $\textcircled{\textsf{S}}$ we obtain (provably in the formal system $\vdash$) that $\langle\varphi\rangle$ is the unique y for which $\textcircled{\textsf{S}}(\langle\chi\rangle, y)$ holds.
- From this last fact and (2), we obtain (1).
A similar argument works for soft self-substitution. □

Prop 6

Proof. Immediately from Prop. 5, taking $\psi(x)$ to be $\neg\ \textcircled{\textsf{T}}(x)$ and $\neg\ (\exists y.\textcircled{\textsf{P}}(y, x) \wedge \mathsf{RosserTwist}(y, x))$, respectively. □

Prop 7

Proof. Let G be a Gödel sentence. To prove $\nvdash \mathsf{G}$, we assume (1) $\vdash \mathsf{G}$ and aim to reach a contradiction.
- Since G is a Gödel sentence, we obtain $\vdash \neg \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$.
- From (1) and HBL_1, we obtain $\vdash \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$.
- The last two facts contradict the consistency assumption. □

Theorem 9

Proof. (1): Immediate from Prop. 6.
(2): $\nvdash \neg \mathsf{G}$ follows by applying Prop. 8 to the assumptions, so it remains to show $\nvdash \mathsf{G}$.
- From OCon, we obtain Con.
- Applying Lemma 24 to $\mathsf{Rel}_{\vdash}^{\Vdash}$ and $\mathsf{Repr}_{\Vdash}$, we obtain HBL_1.
- Applying Prop. 7 to the last to facts, we obtain $\nvdash \mathsf{G}$, as desired. □

Theorem 10

Proof. (1): Immediate from Prop. 6.
(2): $\nvdash \mathsf{R}$ follows by applying Prop. 26 to the assumptions, and $\nvdash \neg \mathsf{R}$ follows by applying Prop. 25 to the assumptions. □

Theorem 11

Proof.
- Applying Lemma 3(2) to the assumptions, we obtain OCon.
- Applying Lemma 4 to the assumptions, we obtain $\mathsf{Rel}_{\vdash}^{\Vdash}$, $\mathsf{Repr}_{\Vdash}$ and $\mathsf{Clean}_{\Vdash}$, where:
— Proof and $\Vdash$ are defined as in Lemma 4 and
— $\Vdash$ is represented by P.
- Applying Theorem 9 to the last facts, OCon and the assumptions, we obtain that:
— there exists a Gödel sentence and
— letting G be a Gödel sentence $\nvdash \mathsf{G}$ and $\nvdash \neg \mathsf{G}$.
It remains to show $\models \mathsf{G}$.
- From $\nvdash \mathsf{G}$, $\mathsf{Rel}_{\vdash}^{\Vdash}$ and $\mathsf{Repr}_{\Vdash}$, we obtain $\vdash \neg \mathsf{P}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$.
- With $\mathsf{Soundness}$, we obtain $\models \neg \mathsf{P}(n, \langle \mathsf{G} \rangle)$ for all $n \in \mathsf{Num}$.
- With $\mathsf{Syn}_{\models}(3)$, we obtain (*) $\models \forall x. \neg \mathsf{P}(x, \langle \mathsf{G} \rangle)$.
- From $\mathsf{Rel}_{\vdash}^{\Vdash}$ and intuitionistic logic, we obtain $\vdash (\forall x. \neg \mathsf{P}(x, \langle \mathsf{G} \rangle)) \rightarrow \neg \vdash \langle \mathsf{G} \rangle$.
- With $\mathsf{Soundness}$, we obtain $\models (\forall x. \neg \mathsf{P}(x, \langle \mathsf{G} \rangle)) \rightarrow \neg \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$.
- With $\mathsf{Syn}_{\models}(2)$ and (*), we obtain $\models \neg \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$.
- With the definition of Gödel sentence and $\mathsf{Soundness}$, we obtain $\models \mathsf{G}$, as desired. □

Theorem 12

Proof. (1): Immediate from Prop. 6.
(2): Let G be a Gödel sentence. We know that $\nvdash \mathsf{G}$ from Prop. 7, so we are left to prove $\nvdash \neg\, \mathsf{G}$. To this end, we assume (*) $\vdash \neg\, \mathsf{G}$ and aim to reach a contradiction.
- Since G is a Gödel sentence, we obtain $\vdash \neg\, \neg\, \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$.
- With classical logic, we obtain $\vdash \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$.
- With $\mathsf{HBL}_1^{\Leftarrow}$, we obtain $\vdash \mathsf{G}$.
- With (*), this contradicts the consistency assumption. □

Theorem 13

Proof. Let G be a Gödel sentence, whose existence is ensured by Prop. 6 and $\mathsf{Repr}_{\mathsf{S}}$; by Prop. 7, Con and HBL_1, we also have $\nvdash \mathsf{G}$, i.e. (by the Gödel sentence definition), (1) $\nvdash \neg \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$.

To prove (2) $\nvdash \neg\, \textcircled{\scriptsize P}\langle \bot \rangle$, it suffices to prove (3) $\vdash \textcircled{\scriptsize P}\langle \mathsf{G} \rangle \rightarrow \textcircled{\scriptsize P}\langle \bot \rangle$. Indeed, the last would imply $\vdash \neg\, \textcircled{\scriptsize P}\langle \bot \rangle \rightarrow \neg\, \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$, which together with (1) would imply (2).

So it remains to prove (3).
- Since G is a Gödel sentence, we obtain $\vdash \mathsf{G} \rightarrow \neg\, \textcircled{\scriptsize P}\langle \mathsf{G} \rangle$.
- With HBL_1, we obtain $\vdash \textcircled{\scriptsize P}\langle \mathsf{G} \rightarrow \neg\, \textcircled{\scriptsize P}\langle \mathsf{G} \rangle \rangle$.
- With HBL_2 we obtain $\vdash \textcircled{\scriptsize P}\langle \mathsf{G} \rangle \rightarrow \textcircled{\scriptsize P}\langle \neg\, \textcircled{\scriptsize P}\langle \mathsf{G} \rangle \rangle$.
- From HBL_3, we obtain $\vdash \textcircled{\scriptsize P}\langle \mathsf{G} \rangle \rightarrow \textcircled{\scriptsize P}\langle \textcircled{\scriptsize P}\langle \mathsf{G} \rangle \rangle$.
- From HBL_2 we obtain $\vdash \textcircled{\scriptsize P}\langle \textcircled{\scriptsize P}\langle \mathsf{G} \rangle \rangle \wedge \textcircled{\scriptsize P}\langle \neg\, \textcircled{\scriptsize P}\langle \mathsf{G} \rangle \rangle \rightarrow \textcircled{\scriptsize P}\langle \bot \rangle$.
- From the last three facts, we obtain (3), as desired. □

Theorem 17

Proof. The only time when (1) $\vdash \mathsf{jcon}$ is used in the proof is via its consequence $\vdash \neg\, (\textcircled{\scriptsize P}\langle \varphi \rangle \wedge \textcircled{\scriptsize P}(\textcircled{\scriptsize N}\langle \varphi \rangle))$, which by $\mathsf{Repr}_{\mathcal{F}}$ and CapN would follow from (**) $\vdash \neg\, (\textcircled{\scriptsize P}\langle \varphi \rangle \wedge \textcircled{\scriptsize P}\langle \neg\, \varphi \rangle)$. So it suffices to show that the last follows from $\vdash \neg\, \textcircled{\scriptsize P}\langle \bot \rangle$, WHBL_2 and HBL_4:
- From HBL_4, we obtain $\vdash \textcircled{\scriptsize P}\langle \varphi \rangle \wedge \textcircled{\scriptsize P}\langle \neg\, \varphi \rangle \rightarrow \textcircled{\scriptsize P}\langle \varphi \wedge \neg\, \varphi \rangle$.
- From WHBL_2 and $\vdash \varphi \wedge \neg\, \varphi \rightarrow \bot$, we obtain $\vdash \textcircled{\scriptsize P}\langle \varphi \wedge \neg\, \varphi \rangle \rightarrow \textcircled{\scriptsize P}\langle \bot \rangle$.
- From the last two facts, we obtain $\vdash \textcircled{\scriptsize P}\langle \varphi \rangle \wedge \textcircled{\scriptsize P}\langle \neg\, \varphi \rangle \rightarrow \textcircled{\scriptsize P}\langle \bot \rangle$.
- Hence $\vdash \neg\, \textcircled{\scriptsize P}\langle \bot \rangle \rightarrow \neg\, (\textcircled{\scriptsize P}\langle \varphi \rangle \wedge \textcircled{\scriptsize P}\langle \neg\, \varphi \rangle)$.
- With $\vdash \neg\, \textcircled{\scriptsize P}\langle \bot \rangle$, we obtain (**), as desired. □