# Incompleteness in a General Setting

John L. Bell

Full proofs of the Gödel incompleteness theorems are highly intricate affairs. Much of the intricacy lies in the details of setting up and checking the properties of a coding system representing the syntax of an object language (typically, that of arithmetic) within that same language. These details are seldom illuminating and tend to obscure the core of the argument. For this reason a number of efforts have been made to present the essentials of the proofs of Gödel's theorems without getting mired in syntactic or computational details. One of the most important of these efforts was made by Löb [8] in connection with his analysis of sentences asserting their own provability. Löb formulated three conditions (now known as the *Hilbert-Bernays-Löb* derivability conditions), on the provability predicate in a formal system which are jointly sufficient to yield the Gödel's second incompleteness theorem for it. A key role in Löb's analysis is played by (a special case of) what later became known as the *diagonalization* or *fixed point* property of formal systems, a property which had already, in essence, been exploited by Gödel in his original proofs of the incompleteness theorems. The fixed point property plays a central role in Lawvere's [7] category-theoretic account of incompleteness phenomena (see also [10]). Incompleteness theorems have also been subjected to intensive investigation within the framework of *modal logic* (see, e.g.[4], [5]). In this formulation the modal operator takes up the role previously played by the provability predicate, and the derivability conditions on the latter are translated into algebraic conditions (the so-called *GL,* i.e., Gödel–Löb, conditions) on the former.

My purpose here is to present a framework for incompleteness phenomena, fully compatible with intuitionistic or constructive principles, in which the idea of a coding system is *retained*, only in a simple, but very general form, a form wholly free of syntactical notions. As codes we shall take the elements of an arbitrary given nonempty set, possibly, but not necessarily, the set of natural numbers. As the objects to be encoded we take the elements of a second arbitrary nonempty set called the set of *sentences*: these are the counterpart of the sentences of a given formal language. We shall also suppose that the set of sentences is equipped with an equivalence relation which corresponds to the relation of provable equivalence with respect to a theory. Equivalence classes with respect to this equivalence relation will be called *propositions*.[1]

We shall take as our background theory intuitionistic set theory in any of its usual formulations (e.g. that presented in [1])[2]. We assume given two sets: $\Sigma$, the set of *sentences*, and $C$, the set of *codes*: we also assume that both $\Sigma$ and $C$ contain at least one element. The elements of the exponential[3] $\Sigma^C$ may be considered as corresponding to formulas with one free variable ranging over $C$. Constant elements of $\Sigma^C$ may be identified with sentences. For $\sigma \in \Sigma$ we write $\sigma$ for the map $C \to \Sigma$ with constant value $\sigma$.

We also assume given:

---

[1] We note that for classical theories propositions in the above sense form a Boolean algebra, but for the intuitionistic theories which we shall have in mind propositions constitute a *Heyting* algebra, that is, a (distributive) lattice $(L, \leq, \wedge, \vee)$ with top and bottom elements 1, 0 equipped with a binary operation $\Rightarrow$ satisfying $x \wedge y \leq z$ iff $x \leq y \Rightarrow z$. We define the operations $\neg$ and $\Leftrightarrow$ by $\neg x = x \Rightarrow 0$ and $x \Leftrightarrow y = (x \Rightarrow y) \wedge (y \Rightarrow x)$. When the elements of a Heyting algebra are regarded as propositions arising from a theory, the relation $\leq$ represents entailment and the operations $\wedge$, $\vee$, $\Rightarrow$, $\neg$, $\Leftrightarrow$, 0, 1 represent conjunction, disjunction, implication, negation, bi-implication, and refutable and provable propositions, respectively. For a proposition $a$, the assertion that $a = 1$ expresses the condition that $a$ is *provable*. The *consistency* of the theory is expressed by the assertion that $0 \neq 1$, that is , by the assertion that its corresponding algebra of propositions has at least two elements.

[2] For definiteness we could take our background theory to be Zermelo set theory formulated within intuitionistic first-order logic. For much of the development in the paper, one may take classical set theory as background theory and note that no use of the law of excluded middle is made.

[3] The exponential $X^Y$ of two sets $X$, $Y$ is the set of functions from $Y$ to $X$.

- An equivalence relation $\approx$ on $\Sigma$. For $\sigma \in \Sigma$ we write $[\sigma]$ for the $\approx$-equivalence class of $\sigma$ and $\Omega(\Sigma)$, or simply $\Omega$, for the set of all such equivalence classes. The members of $\Omega$ are called *propositions*, and for $\sigma \in \Sigma$, $[\![\sigma]\!]$ is called the proposition *associated* with the sentence $\sigma$. Any $\approx$-preserving self-map $f$ on $\Sigma$ induces a self map $f^\dagger$ on $\Omega$ defined by $f^\dagger([\![\sigma]\!]) = [\![ff(\sigma)]\!]$ for $\sigma \in \Sigma$. We sometimes identify a $\approx$-preserving self-map on $\Sigma$ with the self-map on $\Omega$ it induces: such maps may thus be thought of as acting on $\Omega$.
- A submonoid[4] $L$ of $\Sigma^\Sigma$, each member of which preserves $\approx$. The members of $L$ are called *logical maps* (*on* $\Sigma$). The self-maps on $\Omega$ induced by logical self-maps on $\Sigma$ are called *logical maps on* $\Omega$.
- A subset $K$ of $\Sigma^C$ containing each constant function $\sigma$. The members of $K$ are called *coded functions.* We suppose that $L$ acts on $K$, that is, $K$ is closed under composition with members of $L$ on the left.
- A *coding map* $k$: $K \to C$. We shall usually write $\ulcorner\varphi\urcorner$ for $k(\varphi)$. By abuse of notation we shall, for $\sigma \in \Sigma$, write simply $\ulcorner\sigma\urcorner$ for $\ulcorner\sigma\urcorner$. The elements $\ulcorner\varphi\urcorner$, $\ulcorner\sigma\urcorner$ are called the *codes* of $\varphi$ and $\sigma$ respectively.

We shall call a sextuple $\mathrm{A} = (\Sigma, C, \approx, K, L, k,)$ subject to the above data a *coding assemblage.* If $\Omega(\Sigma)$ has at least two elements, $\mathrm{A}$ is said to be *consistent.*

We now assume a fixed coding assemblage $\mathrm{A}$ to be given.

A self-map $f$ on $\Sigma$ is *codable* if, for some map $\varphi \in K$, we have $f(\sigma) \approx \varphi(\ulcorner\sigma\urcorner)$ for all $\sigma \in \Sigma$. In this sense $f$ can be represented, up to $\approx$-equivalence, as a (coded) function of codes. Such a map $\varphi$ is called a *coding representation* for $f$. Given $\varphi \in K$, the map $\varphi^\dagger$: $\sigma \mapsto \varphi(\ulcorner\sigma\urcorner)$: $\Sigma \to \Sigma$ is called the self-map on $\Sigma$ *induced by* $\varphi$; if

[4] That is, a subset of $\Sigma^\Sigma$ containing the identity map on $\Sigma$ and closed under composition.

$\varphi^\dagger$preserves $\approx$, we shall say that $\varphi$ is *equable.* Clearly, if $\varphi$ is equable, $\varphi^\dagger$ is then codable.

Self-maps on $\Omega$ induced by codable $\approx$-preserving self-maps on $\Sigma$ are called *codable self-maps on* $\Omega$.

We can now prove

**Proposition 1.** The following conditions are equivalent:

(i) every logical map on $\Sigma$ is codable;

(ii) the identity map $1_\Sigma$ on $\Sigma$ is codable.

**Proof.** (i) $\Rightarrow$ (ii) is obvious. For the converse, the assumption gives a map $\tau \in K$ for which $\sigma \approx \tau(\ulcorner\sigma\urcorner)$ for all $\sigma \in \Sigma$. Clearly, for any logical map *f*, the map $f \circ \tau$ is a coding representation for *f.* ■

A coding representation $\tau$ for $1_\Sigma$ is called a *Tarski map*: it satisfies

$$\sigma \approx \tau(\ulcorner\sigma\urcorner).$$

This is the counterpart in our framework of what is called in the literature *Tarski's T-scheme*. A Tarski map therefore corresponds to a *truth definition.* From the equation above we see that a Tarski map, or truth definition, is just a *left inverse,* up to $\approx$-equivalence, for the coding map $\sigma \mapsto \ulcorner\sigma\urcorner\colon \Sigma \to C$.

Our next task is to prove a *fixed point lemma.*

Let us call a self-map *f* on $\Sigma$ *diagonalizable* if, for some $\varphi \in K$, we have

$$f(\psi(\ulcorner\psi\urcorner)) \approx \varphi(\ulcorner\psi\urcorner)$$

for all $\psi \in K$. $\varphi$ is called a *diagonal representation* for *f.* Notice that if $f$ is diagonalizable, and $p$ logical, then $p \circ f$ is diagonalizable.

A $\approx$-*fixed point* for a self map $f$ on $\Sigma$ is an element $\sigma \in \Sigma$ such that $f(\sigma) \approx \sigma$.

**Lemma 1.**

(i) Any diagonalizable self-map on $\Sigma$ has a $\approx$-fixed point.

(ii) The composite of a diagonalizable $\approx$-preserving map with a logical map has a $\approx$-fixed point.

**Proof.** (i) If $f$ is a diagonalizable self-map on $\Sigma$ with diagonal representation $\varphi: C \to \Sigma$, then $\varphi(\ulcorner\varphi\urcorner)$ is easily seen to be a $\approx$-fixed point for $f$.

(ii) Let $f$ be a diagonalizable $\approx$-preserving self-map on $\Sigma$ and $p$ a logical map. Then, as observed above, $p \circ f$ is diagonalizable and so, by (i), has a $\approx$-fixed point $\sigma$. Then $f(p(f(\sigma))) \approx f(\sigma)$, so that $f(\sigma)$ is a $\approx$- fixed point for $f \circ p$. ■

A *diagonal map* for $\mathrm{A}$ is a map $d: C \to C$ such that

(i) $d(\ulcorner\varphi\urcorner) = \ulcorner\varphi(\ulcorner\varphi\urcorner)\urcorner$ for $\varphi \in K$;

(ii) $K$ is closed under composition with $d$ on the right: if $\varphi \in K$, then $\varphi \circ d \in K$.

**Lemma 2.** If $\mathrm{A}$ has a diagonal map, then every codable self-map on $\Sigma$ is diagonalizable.

**Proof.** Given a codable self -map $f$ on $\Sigma$ with coding representation $\varphi \in K$, let $\varphi^* = \varphi \circ d \in K$. Then for $\psi \in K$, we have

$$f(\psi(\ulcorner\psi\urcorner)) \approx \varphi(\ulcorner\psi(\ulcorner\psi\urcorner)\urcorner) \approx \varphi(d(\ulcorner\psi\urcorner)) \approx \varphi^*(\ulcorner\psi\urcorner).$$

So $\varphi^*$ is a diagonal representation for $f$, and the latter is accordingly diagonalizable. ■

Lemmas 1 and 2 immediately yield the

**Fixed Point Lemma.** Suppose that the coding assemblage $\mathrm{A}$ has a diagonal map. Then:

(i) Any codable self-map on $\Sigma$ has a $\approx$-fixed point.

(ii) The composite of a $\approx$-preserving self-map on $\Sigma$ with a logical map has a $\approx$-fixed point.

(iii) Any codable self-map on $\Omega$, as well as its composite with a logical map, has a fixed point. ■

It now follows from Proposition 1 and the Fixed Point Lemma that if a coding assemblage has both a diagonal map and a Tarski map, then every logical map on $\Sigma$ has a $\approx$-fixed point. This immediately yields

**Tarski's Theorem.** Suppose that $\mathrm{A}$ has a diagonal map and a logical map with no $\approx$-fixed points (equivalently, if $\Omega$ has a logical map with no fixed points), then $\mathrm{A}$ has no Tarski map. ■

Tarski's theorem in this formulation applies in particular when $\Omega$ is a Heyting algebra with at least two elements, and the negation map $\neg$ on $\Omega$—which then has no fixed points—is included among the logical maps. This in turn enables us to recapture the usual formulation of Tarski's theorem on the undefinability of truth. For the algebra of propositions of a consistent theory $T$ has at least two elements, so, provided the language of $T$ meets the modest requirements for generating a coding assemblage (along the lines of example 1 immediately below), it follows that the associated coding assemblage has no Tarski map. This means that the language for $T$ contains no truth definition for $T$; in a word, truth for $T$ is *undefinable* in $T$.

**Examples**

**1. Peano arithmetic.** In this case the ingredients of the coding assemblage $\mathcal{P}$ — the *Peano assemblage* – are as follows: $\Sigma$ is the set of sentences of the language $\mathcal{L}$ of first-order intuitionistic arithmetic **P**, $C$ is the set $N$ of natural numbers, $\approx$ is the relation of provable equivalence from **P**[5] $K$ is the set of maps of the form $\varphi$: $n \mapsto \psi(\mathbf{n})$ where $\psi(x)$ is a formula of $\mathcal{L}$ with at most one free variable and $\mathbf{n}$ is the term of $\mathcal{L}$ representing $n$. The coding map $k$ is given by $k(\varphi) = \#\psi$ where # is any standard Gödel numbering of the formulas of $\mathcal{L}$. Finally $L$ consists of the maps $\sigma \mapsto \sigma$, $\sigma \mapsto \neg\sigma$, $\sigma \mapsto \neg\neg\sigma$. $\mathcal{P}$ has a diagonal map $d$: $N \to N$ given by setting $d(m) = s(m,m)$, where $s$: $N \times N \to N$ is a recursive substitution function on Gödel numbers[6].

Assuming that $\mathcal{P}$ is consistent, it follows, as observed above, that $\mathcal{P}$ has no Tarski map, and so truth in $\mathcal{P}$ is undefinable in $\mathcal{P}$.

**2. Intuitionistic set theory.** Just as in classical set theory the power set P$A$ of any set $A$ is a Boolean algebra under the usual set-theoretic operations, so in intuitionistic set theory the power set is, under the same operations, a Heyting algebra. In particular, writing 1 for the one-element set {0}, P1 is a Heyting algebra which we shall denote by $\prod$. If $\sigma$ is a sentence of the language of set theory, we write $\{0 \mid \sigma\}$ for the element $\{x\colon x = 0 \wedge \sigma\}$ of $\prod$. From the axiom of extensionality it follows that $\{0 \mid \sigma\} = \{0 \mid \sigma'\}$ iff $\sigma \leftrightarrow \sigma'$. Thus the elements of $\prod$ correspond naturally to what we have termed *propositions,* in this case, to sentences identified under provable equivalence from the axioms of intuitionistic

[5] Thus $\Omega$ may be regarded as the set of sentences of $\mathcal{L}$ identified up to provable equivalence from **P.**

[6] See, e.g. [3], Example 7.4.5.

set theory. Under this correspondence each element $\omega \in \prod$ is correlated with the proposition $0 \in \omega$, and each proposition $\sigma$ with the element $\{0 \mid \sigma\}$ of $\prod$.

$\prod$ also plays the role of a *subset classifier.* That is, for each set $A$, subsets of $A$ are correlated bijectively with maps $A \to \prod$: each subset $X \subseteq A$ is correlated with the map $x \mapsto \{0 \mid x \in X\} : A \to \prod$, and each map $f : A \to \prod$ with the subset $f^{-1}(1)$ of $A$. The top element 1 (bottom element $\varnothing$) 1 of $\prod$ is identified with the true (false) proposition(s). In this way $\prod^A$ is seen to be naturally isomorphic to P$A$.

Now let us attempt to build a coding assemblage using $\prod$ as the underlying set of sentences and the identity relation as the underlying equivalence relation. Here it is natural to take $C$, the set of codes, to be any set containing at least one element, and to take $L = \prod^{\prod}$ and $K = \prod^C$. Using the observation immediately above, we may then identify $K$ with P$C$. Take the coding map $k$ to be an arbitrary map P$C \to C$; for $X \in$ P$C$, write $\ulcorner X \urcorner$ for $k(X)$.

For $\omega \in \prod$, the constant map $\omega : C \to \prod$ is correlated with the element $\omega^*$ $= \{x \in C : 0 \in \omega\}$ of P$C$; it will be convenient to write $\ulcorner\omega\urcorner$ for $\ulcorner\omega^*\urcorner$. The map $\omega \mapsto \ulcorner\omega\urcorner$: $\prod \to C$ is the *coding map* on $\prod$.

The sextuple $\mathrm{Q} = (\prod, C, =, \mathrm{P}C, \prod^{\prod}, k)$ is accordingly a coding assemblage. Does $\mathrm{Q}$ have a diagonal map? As we shall see, this cannot be done when the coding map on $\prod$ satisfies the modest requirement of being injective[7].

In fact, if $\ulcorner\bullet\urcorner$ is injective, a diagonal map $d$ would then have to satisfy

$$(*) \qquad d(\ulcorner X \urcorner) = \ulcorner\{0 \mid \ulcorner X \urcorner \in X\}\urcorner$$

for $X \in$ P$C$. Now define

[7] The modesty of this requirement is more easily seen when the background theory is classical (i.e. the law of excluded middle holds). For then $\Omega = \{\varnothing, 1\}$ and injectivity of $\ulcorner\bullet\urcorner$ boils down simply to $\ulcorner\varnothing\urcorner \neq \ulcorner 1 \urcorner$: that is, the true and the false receive different codes.

$$U = \{x \in C \colon d(x) = \ulcorner\varnothing\urcorner\}.$$

Then, using (*) and the injectivity of $\ulcorner\bullet\urcorner$,

$$\begin{aligned} \ulcorner U \urcorner \in U \;\leftrightarrow\; & d(U) = \ulcorner\varnothing\urcorner \\ \leftrightarrow\; & \ulcorner\{0 \mid \ulcorner U \urcorner \in U\}\urcorner = \ulcorner\varnothing\urcorner \\ \leftrightarrow\; & \{0 \mid \ulcorner U \urcorner \in U\} = \varnothing \\ \leftrightarrow\; & \ulcorner U \urcorner \notin U \end{aligned}$$

and we have a contradiction.

We now turn to Gödel's theorems. *Henceforth we shall assume that the coding assemblage* $\mathrm{A}$ *has a diagonal map* $d$.

We have seen that every codable self-map $f$ on $\Sigma$ has a $\approx$-fixed point. Let us call an element $\alpha \in \Sigma$ a *strong* $\approx$-fixed point for $f$ if, for all $\sigma \in \Sigma$, we have

(*) $$f(\sigma) \approx \alpha \leftrightarrow \sigma \approx \alpha.$$

We next prove another version of the Fixed Point Lemma, namely, the

**Strong Fixed Point Lemma.** Suppose that $\Sigma$ has a codable $\approx$-preserving self-map $f$ with a strong $\approx$-fixed point $\alpha$. Then, for any logical map $p$ on $\Sigma$ there is $\beta \in \Sigma$ such that

$$p(\beta) \approx \alpha \leftrightarrow \beta \approx \alpha.$$

**Proof.** By the Fixed Point Lemma, $f \circ p$ has a $\approx$-fixed point $\beta$. We then have, using (*),

$$p(\beta) \approx \alpha \leftrightarrow f(p((\beta)) \approx \alpha \leftrightarrow \beta \approx \alpha. \qquad \blacksquare$$

Now we can formulate Gödel's First Incompleteness Theorem in the present setting. Here we require $\Sigma$ to have a distinguished element $\top$: we think of $\top$ as representing the *provable* sentences in the sense that the provable sentences are taken to be precisely those $\approx$--equivalent to $\top$. We suppose given an equable map $\pi \in K$ which we shall term a *provability* map, in the sense that, for each $\sigma \in \Sigma$, the element $\pi(\ulcorner\sigma\urcorner)$ of $\Sigma$ shall be construed as the sentence $\sigma$ *is provable.* The self-map $g$ on $\Sigma$ induced by $\pi$ is then necessarily codable (as well as $\approx$-preserving). We shall call $g$ a *Gödel map* if it has $\top$ as a strong $\approx$-fixed point. If $g$ is a Gödel map, then the provability map $\pi$ satisfies

$$\sigma \approx \top \leftrightarrow \pi(\ulcorner\sigma\urcorner) \approx \top.$$

This may be construed as asserting that a sentence $\sigma$ is provable iff the sentence $\pi(\ulcorner\sigma\urcorner)$ asserting the provability of $\sigma$ is itself provable. Notice that the self-map on $\Sigma$ induced by a Tarski map is a Gödel map.

Now call a coding assemblage *Gödelian* if it is consistent, has a Gödel map, and there is an element $\perp$ of $\Sigma$ such that $\top \not\approx \perp$ together with a logical map $\nu$ on $\Sigma$ such that $\nu(\top) \approx \perp$ and $\nu(\perp) \approx \top$.

We think of $\nu$ as the *negation* operation on sentences and $\perp$ as representing the *refutable* sentences.

We can now prove

**Gödel's First Incompleteness Theorem.** The set of propositions of any Gödelian coding assemblage has at least three elements.

**Proof.** Given a Gödelian coding assemblage there is, by the Strong Fixed Point Lemma, an element $\beta \in \Sigma$ for which $\nu(\beta) \approx \top \leftrightarrow \beta \approx \top$. In that case $\top \not\approx \beta \not\approx \perp$, and thus $\Omega$ has the three distinct elements $[\![\top]\!]$, $[\![\perp]\!]$, $[\![\beta]\!]$. ■

An element $\beta \in \Sigma$ such that $\top \not\approx \beta \not\approx \bot$ evidently represents an *undecidable* sentence, so the theorem just proved may be taken to assert that *any Gödelian coding assemblage contains undecidable sentences.*

All this applies in particular to the Peano coding assemblage $\mathrm{P}$. Let $\top$ be the sentence $\mathbf{0} = \mathbf{0}$, and $\bot$ the sentence $\mathbf{0} = \mathbf{1}$. Also let *Prov* be a provability predicate for **P.** Then, by standard arguments[8], we have, for any arithmetical sentences $\alpha$, $\beta$,

(**Prov$_1$**) $\vdash_{\mathbf{P}} \alpha \leftrightarrow \vdash_{\mathbf{P}} Prov(\#\alpha)$

(**Prov$_2$**) $\vdash_{\mathbf{P}} Prov(\#(\alpha \to \beta)) \to [Prov(\#\alpha) \to Prov(\#\beta)])$

(**Prov$_3$**) $\vdash_{\mathbf{P}} Prov(\#\alpha) \to Prov(Prov(\#\alpha))$.

Now let $\pi$: $N \to \Sigma$ be the map $n \mapsto Prov(\mathbf{n})$. It follows from (**Prov$_2$**) that $\pi$ is equable, and from (**Prov$_1$**) that the map $g$: $\Sigma \to \Sigma$ induced by $\pi$ has $\top$ as a strong $\approx$-fixed point. Accordingly $g$ is is a Gödel map for $\mathrm{P}$. Assuming that **P** is consistent, $\mathrm{P}$ is then Gödelian, and accordingly contains undecidable propositions.

Finally let us set about formulating Gödel's Second Incompleteness Theorem in the present setting. To do this we need to introduce the concept of a Hilbert-Bernays-Löb, or HBL-operator. Let us assume that $\mathrm{A}$ is a coding assemblage in which $\Omega$ is a Heyting algebra[9]. An *HBL-operator* in $\mathrm{A}$ is a codable self-map $\square$ on $\Omega$ satisfying the conditions:

---

[8] See, e.g. [6], Ch. 16.

[9] We think of $\Omega$`s top element 1, and bottom element 0 as representing, respectively, the *provable* and *refutable* sentences.

(a) $\Box 1 = 1$

(b) $\Box(x \Rightarrow y) \leq (\Box x \Rightarrow \Box y)$

(c) $\Box x \leq \Box\Box x$.

An HBL-operator may be considered a modal operator satisfying the K4 axioms[10]. It follows quite easily from (a) and (b) that $\Box$ preserves $\wedge$ , and hence is also order-preserving.

We may think of $\Box$ as a provability operator acting on propositions: for each proposition $x$, $\Box x$ is the proposition asserting "$x$ is provable". In that case (a) above asserts: *if x is a provable proposition, then so is the proposition "x is provable"*; (b) asserts: *the proposition "x implies y is provable" implies the proposition '"x is provable" implies "y is provable"'*; and (c) asserts: *the proposition "x is provable" implies the proposition "'x is provable' is provable".*

Now let us call a coding assemblage A *suitable*[11] if (i) $\Omega$ is a Heyting algebra with an HBL operator and (ii) for each $a \in \Omega$ the map $x \mapsto (\Box x \Rightarrow a)$: $\Omega \to \Omega$ is codable.

We can now prove a version of

**Löb's Theorem**[12]**.** Let A be a suitable coding assemblage with HBL-operator $\Box$. Then, for any $a \in \Omega$

(i) $\Box(\Box a \Rightarrow a) \leq \Box a$.

(ii) $\Box a \leq a \;\rightarrow\; a = 1$.

[10] See [4], p. 5.
[11] I.e., suitable for proving Gödel's second incompleteness theorem: see below.
[12] Theorem 4.1.1 of [6].

**Proof.** Given $a \in \Omega$, the map $x \mapsto (\Box x \Rightarrow a)\colon \Omega \to \Omega$ is codable and so by the Fixed Point Lemma has a fixed point $b$. That is,

$$(*) \qquad b = (\Box b \Rightarrow a).$$

*A fortiori* $b \leq (\Box b \Rightarrow a)$, whence

$$\Box b \leq \Box(\Box b \Rightarrow a) \leq (\Box\Box b \Rightarrow \Box b).$$

Hence

$$(**) \qquad \Box b = \Box b \wedge \Box\Box b \leq \Box a$$

It follows that $(\Box a \Rightarrow a) \leq (\Box b \Rightarrow a)$, whence

$$\Box(\Box a \Rightarrow a) \leq \Box(\Box b \Rightarrow a)$$

$$= \Box b \ \text{(by (*))}$$

$$\leq \Box a \ \text{(by (**))}.$$

This gives (i).

For (ii), we assume $\Box a \leq a$, so that $(\Box a \Rightarrow a) = 1$. It now follows from (i) that

$$1 = \Box 1 = \Box(\Box a \Rightarrow a) \leq \Box a.$$

Therefore $\Box a = 1$, and since $\Box a \leq a$, we conclude that $a = 1$. ■

**Corollary 1.** Let A be a suitable coding assemblage with HBL-operator $\Box$. Then the following conditions are equivalent (a) $\forall x(x \leq \neg\neg\Box x)$;

(b) $\forall x(\neg\Box x = 0)$; (c) $\neg\Box 0 = 0$. *A fortiori* $\forall x(x \leq \Box x)$ implies $\neg\Box 0 = 0$.

**Proof.** (a) $\rightarrow$ (b). Suppose that $\forall x(x \leq \neg\neg\Box x)$ . Then

$$(*) \qquad \neg\Box x \leq \neg\neg\Box\neg\Box x$$

and by Löb`s Theorem $\Box\neg\Box x = \Box(\Box x \Rightarrow 0) \leq \Box x$. Hence

(**) $$\neg\Box x \leq \neg\Box\neg\Box x.$$

From (*) and (**) we get

$$\neg\Box x \leq \neg\Box\neg\Box x \wedge \neg\neg\Box\neg\Box x = 0.$$

(b) $\rightarrow$ (c) is trivial.

(c) $\rightarrow$ (a). Suppose that $\neg\Box 0 = 0$. Then, for any $x$, $1 = \neg\neg\Box 0 \leq \neg\neg\Box x$, so that $0 = \neg\neg\neg\Box x = \neg\Box x$. ■

**Corollary 2.** Let A be a suitable coding assemblage with HBL-operator $\Box$. Then $\neg\Box 0$ is the unique fixed point of the map $x \mapsto \neg\Box x$.

**Proof.** By Löb's Theorem $\Box\neg\Box 0 = \Box(\Box x \Rightarrow 0) \leq \Box 0$, so that

(*) $$\neg\Box 0 \leq \neg\Box\neg\Box 0.$$

On the other hand $\Box 0 \leq \Box\neg\Box 0$ so that

(**) $$\neg\Box\neg\Box 0 \leq \neg\Box 0.$$

(*) and (**) give $\neg\Box\neg\Box 0 = \neg\Box 0$, and so is $\neg\Box 0$ a fixed point of the map $x \mapsto \neg\Box x$.

To see that $\neg\Box 0$ is the only fixed point, suppose that $a = \neg\Box a$. Then

(***) $$a = \neg\Box a \leq \neg\Box 0.$$

Also $\Box a \leq \neg\neg\Box a = \neg a$, so that $\Box a \leq \Box\Box a \leq \Box\neg a$, whence

$$\Box a \leq \Box a \wedge \Box\neg a = \Box(a \wedge \neg a) = \Box 0.$$

Therefore $\neg\Box 0 \leq \neg\Box a$, so that, by (***), $a = \neg\Box a = \neg\Box 0$. ■

From (i) of Löb's Theorem we see that $\Box$ satisfies the so-called *GL* (Gödel-Löb) *axiom*[13] for a normal modal logic, i.e. the scheme

$$\Box(\Box A \to A) \to \Box A.$$

From Löb's Theorem one derives:

**Gödel's Second Incompleteness Theorem.** Given a suitable consistent coding assemblage A with HBL operator $\Box$. Then, for any $x \in \Omega$, $\Box x \neq 0$; or equivalently, $\neg\Box x \neq 1$. In particular, $\Box 0 \neq 0$; or equivalently $\neg\Box 0 \neq 1$.

**Proof.** If $\Box x = 0$, then $\Box 0 \leq \Box x = 0$; hence by Löb's theorem $0 = 1$, and it follows that A is inconsistent. ■

Consider again the Peano assemblage P. There $\Omega$ is a Heyting algebra and conditions **Prov$_{1\text{-}3}$** on the provability predicate imply that the self-map $\Box$ on $\Omega$ induced by the Gödel map $g$ is an HBL-operator.[14] It is also easily checked for each $a \in \Omega$ the map $x \mapsto (\Box x \Rightarrow a)$: $\Omega \to \Omega$ is codable. Accordingly P is suitable, and so the 2nd incompleteness theorem applies to it.

If we think of $\Box$ as a provability operator, $\neg\Box x$ is the proposition "*x is unprovable*", so that $\neg\Box x = 1$ may be taken as asserting the provability of "*x is unprovable".* In that case the second incompleteness theorem, may be taken to assert that in any suitable consistent coding assemblage, *there is no proposition whose unprovability is provable.* This appies, in particular, to the proposition 0, so that *it is unprovable that* "0 *is unprovable*" Now "0 *is unprovable*" means "*no refutable proposition is also provable*", and it is natural to paraphrase this as " A *is internally consistent*". This terminology enables the second incompleteness theorem as stated above to assume a more familiar form: in any suitable

[13] See [5], p. 5.

[14] **Prov$_1$** actually asserts the stronger condition $\Box x = 1 \leftrightarrow x = 1$, which does not hold for HBL-operators in general.

consistent coding assemblage, *its internal consistency is unprovable.* This applies in particular to the Peano assemblage.

The idea of internal consistency can be extended to the following concordance:

| Proposition | Paraphrase |
|---|---|
| $\Box 0$ | A *is internally inconsistent* |
| $\neg\Box 0$ | A *is internally consistent* |
| $\neg\neg\Box 0$ | A *is weakly internally inconsistent* |
| $\Box\neg\Box 0$ | A *is provably internally consistent* |
| $\neg\Box\neg\Box 0$ | A *is not provably internally consistent* |

In each case, the claim that the proposition is equal to the top element 1 of $\Omega$ is correlated with an assertion about A : for example, $\Box 0 = 1$ with the assertion "A is internally inconsistent" and similarly for the others.

In this spirit, consider (*) of Corollary 2, namely the inequality $\neg\Box 0 \leq \neg\Box\neg\Box 0$. This is equivalent to $(\neg\Box 0 \Rightarrow \neg\Box\neg\Box 0) = 1$, which may be paraphrased: in A, internal consistency implies the unprovability of internal consistency . This is an *internal version* of Gödel's Second Incompleteness Theorem.

In this same spirit, Corollary 2 itself may be translated as: A *is internally consistent* is the unique proposition equivalent to the assertion of its own unprovability. And (he last claim) of Corollary 1 translates as: if, in A, every proposition implies its own provability, then A is weakly internally inconsistent.

Finally, we observe that *consistency and internal inconsistency are compatible.* This follows from the fact that the by the HBL-operator $\Box$ can be taken to be identically 1 - in other words, every proposition can be taken to satisfy the internal condition "__ *is provable*" . All this shows is that internal consistency need have little to do with consistency, or, more generally, that provability maps need have little to do with provability[15].

In conclusion, it should be pointed out that while in stating and proving these results we have used ordinary set-theoretic language, they can be formulated in toposes (see, e.g. [2]) or more general categories (cf. the discussion in [10]).

## References

[1] Bell, John L. *Set Theory: Boolean-Valued Models and Independence Proofs.* Clarendon Press, Oxford, 2005.

[2] Bell, John L. *Toposes and Local Set Theories.* Clarendon Press, Oxford, 1988.

[3] Bell, John L., and Machover, M. *A Course in Mathematical Logic.* North-Holland, 1977.

[4] Boolos, G. *Gödel's second incompleteness theorem explained in words of one syllable.* Mind 103 (1994), no. 409, 1-3.

[5] Boolos, G. *The Logic of Provability.* Cambridge University Press 1995.

[6] Boolos, G. and Jeffrey, R. *Computability and Logic.* Cambridge University Press, 1974.

[7] Lawvere, F.W. *Diagonal arguments and cartesian closed categories.* Category theory, homology theory and their applications, II (Battelle Institute Conference, Seattle, Wash., 1968). Springer, Berlin, 1969, 134-145.

[8] Löb, M. H. *Solution of a problem of Leon Henkin.* Journal of Symbolic Logic 20 (1955), 115-18.

[9] Smorynski, C. *The Incompleteness Theorems.* Handbook of Mathematical Logic, J. Barwise, ed., North-Holland, 1977, pp. 821-866.

[10] Yanofsky, N. *A universal approach to self-referential paradoxes, incompleteness and fixed points.* Bull. Symb. Logic 9(3), 2003, 362-386.

[15] An observation also made in [6].

Department of Philosophy,
University of Western Ontario
e-mail: jbell@uwo.ca