11

Epistemic Church's Thesis and Absolute Undecidability

MARIANNA ANTONUTTI MARFORI AND LEON HORSTEN

11.1 Introduction

On December 26, 1951, Gödel delivered the 25th J. W. Gibbs Lecture at a Meeting of the American Mathematical Association at Brown University. In the lecture, he formulated a disjunctive thesis concerning the limits of mathematical reasoning and the possibility of the existence of mathematical truths that are inaccessible to the human mind. This thesis, known as *Gödel's disjunction*, is introduced as a direct consequence of the incompleteness theorems [Gödel 1951, p. 310]:

> Either ... the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems [henceforth, *absolutely undecidable mathematical sentences*] ... (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives).

That is, either the output of the human mathematical mind exceeds the output of a Turing machine (called the *anti-mechanist thesis*) or there are true mathematical sentences that are undecidable "not just within some particular axiomatic system, but by any mathematical proof the human mind can conceive." The latter are called *absolutely undecidable mathematical sentences*, i.e. mathematical sentences that cannot be either absolutely proved or refuted [Gödel 1951].

According to Gödel, the fact that the disjunctive thesis above holds is a "mathematically established fact" of great philosophical interest which follows from the incompleteness theorems, and as such, it is "entirely independent from the standpoint taken toward the foundation of mathematics" [Gödel 1951]. Indeed, most commentators agree that Gödel's arguments for this disjunctive thesis are compelling.

Since Gödel's disjunction was first formulated in 1951, much effort has gone into finding equally compelling arguments for or against either of the disjuncts. In particular, attempts

Gödel's Disjunction. First Edition. Leon Horsten & Philip Welch (eds).

were made to establish the first disjunct by arguing on *a priori* grounds that the capacities of the human mathematical mind exceed the output of any Turing machine as a consequence of the incompleteness theorems (see chiefly [Lucas 1961], [Penrose 1989], and [Penrose 1994]). These attempts have sparked a lively debate, but there is wide consensus that they have so far turned out to be inconclusive. Fewer efforts have been concentrated on shedding light on the second disjunct, and, at present, no conclusive argument that decides either of the disjuncts has been found.[1] In this chapter, we focus on the second disjunct.

We make use of the framework of *Epistemic Arithmetic* proposed by Shapiro in the mid 1980s. In this framework, an absolute or informal notion of provability is taken as primitive and axiomatically investigated. Since this framework can also express classical (and constructive) mathematical propositions, it constitutes a good setting for the investigation of the concept of absolute undecidability. In this framework, a variant of the Church–Turing thesis can be formulated: this variant has been labelled *ECT* ("Epistemic Church's Thesis") in the literature. While there are strong reasons to think that this variant is not a very faithful approximation of the content of the original Church–Turing thesis, we will show that an analogue of Gödel's disjunction can be established which states that either *ECT* fails, or there are absolutely undecidable propositions (or both).

In analogy with Gödel's disjunction, this raises the question of what the truth value of each of the disjuncts is. While it can be shown that *ECT* implies the existence of absolutely undecidable non-contingent propositions (of low arithmetical complexity), we will see that it seems hard to argue convincingly that *ECT* is true. Consequently, it is not easy to see how the analogue of Gödel's disjunction can be used to show that there are absolutely undecidable propositions. Thus we conclude that the truth value of the disjuncts is not easy to ascertain.

11.2 Absolute Undecidability

11.2.1 Absolute Undecidability in Epistemic Arithmetic

Following suggestions in [Myhill 1960], the notion of informal or absolute provability, and more generally the notion of *a priori* knowability of mathematical propositions, can be investigated in an axiomatic way. In particular, the investigation can be carried out within the framework of epistemic arithmetic developed in [Shapiro 1985b]. This is the framework that shall be adopted throughout the chapter. Since its background is a formal theory of *arithmetic*, the propositions that contain set-theoretic concepts fall outside the scope of the present chapter.

The formal framework of epistemic arithmetic can be described as follows. The formal language $\mathcal{L}_{EA}$ consists of the first-order language of arithmetic plus an intensional propositional operator $\Box$; the arithmetical vocabulary receives its intended interpretation, and the operator $\Box$ is interpreted as *a priori* knowability. The axiomatic theory *EA* that is proposed by Shapiro as describing the laws of *a priori* knowability consists of the axioms of Peano arithmetic plus the laws of **S4** modal logic. It should be noted that the modal logic

S4 contains the necessitation rule and the axiom $\Box\phi \rightarrow \Box\Box\phi$ (the so-called "4 axiom"), so that $\Box$ is indeed an iterable notion.

The *absolute undecidability* of a sentence ϕ can then be expressed as

$$\neg\Box\phi \wedge \neg\Box\neg\phi.$$

This is the notion of absolute undecidability that will figure in the new disjunctive thesis that will be investigated in Section 11.3.

One may wonder whether there are *arithmetical* sentences that are absolutely undecidable in this sense. This is a version of Gödel's question whether there are absolutely undecidable *mathematical* propositions. But one may also ask in this framework whether there are sentences $\phi \in \mathcal{L}_{EA}$ that are undecidable. This question is of course closely related to Gödel's question. But it is not identical to it, for $\mathcal{L}_{EA}$ not only contains mathematical concepts, but also contains the concept of *a priori* knowability, which is not a mathematical concept. Both of these questions about absolute undecidability will be discussed in this chapter.

11.2.2 Other Concepts of Absolute Undecidability

The notion of absolute undecidability in which we are interested here concerns statements that are non-contingent, may contain the concept of absolute or informal provability, and have a determinate truth value.

11.2.2.1 Fitch's Undecidables

In [Fitch 1963], Fitch has argued for the following claim:

Thesis 1 (Fitch) *If there are unknown truths, then there are unknowable truths.*

The kinds of truths that Fitch adduces as witnesses of the consequent of Gödel's disjunction are propositions of the form "p and it is not known that p"; Fitch's argument of course does not provide a concrete witness.

Some believe that Fitch's argument is sound and that furthermore, since the antecedent of its conclusion is true, we must accept that there are unknowable truths. Others think that Fitch's argument is unsound.[2] In any case, Gödel would probably not have been satisfied with Fitch's cases of absolutely undecidable propositions as a way of sharpening Gödel's disjunction (if Fitch's propositions do indeed qualify as absolutely undecidable), for they are *contingent* propositions. And they also fall outside the scope of our analogue of Gödel's disjunction because, using the notion of a priori knowability and arithmetical notions, only non-contingent propositions can be formed in $\mathcal{L}_{EA}$.

11.2.2.2 Formally Undecidable Arithmetical Statements

As said earlier, in his discussion of his disjunctive thesis, Gödel seems to have had *mathematical* undecidable statements in mind. Feferman and Solovay have produced instances of arithmetical sentences of which they claimed that it is unlikely that they will ever be

decided [Feferman and Solovay 1990, Remark 3, p. 292]. However, it is far from clear that the reasons for thinking that such statements will presumably never be decided go as far as establishing that they are *in principle* humanly undecidable, i.e. absolutely undecidable.

Similarly, Boolos has shown that there are infinitely many what he calls *extremely unprovable* arithmetical sentences [Boolos 1982]. These are (true) arithmetical sentences that are not only undecidable in Peano arithmetic, but that are such that Peano arithmetic can only prove them to have properties characterisable in terms of "provability in Peano arithmetic" that *every* arithmetical sentence can be proved (in Peano arithmetic) to have. But, again, Boolos' considerations do not establish that some such sentences cannot be proved *in principle* rather than just in Peano arithmetic. The uncertainty in this area is caused by the fact that we do not have a sufficiently strong grasp on what the right idealisations involved in the notion of absolute undecidability are.[3]

Indeed, it seems that the only argument that we have for establishing of a given arithmetical sentence that it is absolutely unprovable relies essentially on the connection between absolute proof and truth. If we have a proof that ϕ is false (for ϕ arithmetical), then we have *a priori* knowledge that it cannot be provable. Our tentative claim is that this line of argument is the *only* way in which an arithmetical sentence can be shown to be absolutely unprovable. In other words:[4]

Thesis 2 *Provable unprovability of an arithmetical fact supervenes on the provability of the negation of that arithmetical fact.*

The content of this principle can be expressed in the framework of *EA*:

Axiom 1 $\Box\neg\Box\phi \rightarrow \Box\neg\phi$, *for* ϕ *any sentence of the language of arithmetic.*

It is easy to see that:

Proposition 1 *If axiom 1 is true, then there are no provably absolutely undecidable arithmetical sentences.*

The statement that there are no provably absolutely undecidable arithmetical sentences is a restricted version of the **S4.1** axiom (also known as *McKinsey's axiom*) of modal logic.

So, in a nutshell, it seems to us unlikely that we can find a conclusive *a priori* argument that there are absolutely undecidable *arithmetical* sentences. Or, in other words, we think that the second disjunct of Gödel's disjunction *for arithmetical sentences* cannot be established by *a priori* means.

This, however, completely leaves open the question of whether it can be established that there are sentences *of the language of epistemic arithmetic*—i.e. sentences that involve the notion of absolute provability itself—that are absolutely undecidable. Indeed, we cannot appeal to the generalisation of axiom 1 to argue that there are no provably absolutely undecidables in $\mathcal{L}_{EA}$. For if the principle $\Box\neg\Box\phi \rightarrow \Box\neg\phi$ is true for every $\phi \in \mathcal{L}_{EA}$, then absolute provability (demonstrably) coincides with truth. It is clear that for any ϕ, $\Box\neg\Box(\phi \wedge \neg\Box\phi)$. The generalisation of axiom 1 then allows us to conclude $\Box(\phi \rightarrow \Box\phi)$

[Horsten 1997]. In the following sections we will return to the question whether it might be provable in principle that there are absolutely undecidable sentences in $\mathcal{L}_{EA}$.

11.2.2.3 Truth-Indeterminate Undecidables

Another question is whether there are absolutely undecidable sentences that, in addition to the arithmetical vocabulary, contain a primitive notion of provability. Reinhardt has observed that absolutely undecidable sentences can be shown to exist if we have a provably sound *absolute provability predicate* [Reinhardt 1986]:

Proposition 2 *Suppose that $P(x)$ is any formula with x free, and let $S \supseteq EA$ be such that $S \vdash \Box(P(\ulcorner\phi\urcorner) \rightarrow \phi)$. Then there is a sentence G_S such that $S \vdash \Box G_S \wedge \Box\neg P(\ulcorner G_S\urcorner)$.*

If $P(x)$ is an absolute provability predicate satisfying the condition of the proposition, then G_S is an absolutely undecidable sentence.[5]

The sentence G_S is produced by diagonalisation (the fixed-point lemma). Intuitively, G_S is a sentence which says of itself that it is absolutely unprovable. It is not purely arithmetical, due to the predicate P, which figures in the instance of the diagonal lemma that is used to produce G_S.

In contrast to Fitch's propositions, G_S is not contingent: if it is true (or false), then it is so necessarily. So Reinhardt's proposition might be taken to be more relevant to the second disjunct of Gödel's disjunction. However, G_S is the so-called "knower sentence", see [Anderson 1983]. It is a paradoxical sentence: intuitively, it lacks a truth value just like the liar sentence does, and for similar reasons. And if G_S lacks a truth value, then it is not even a candidate for being proven or refuted, so it does not seem very relevant to the assessment of Gödel's disjunction.

There is a rich literature about purely mathematical sentences—mostly set-theoretical—that *may* be truth-indeterminate.[6] The continuum hypothesis is perhaps the most famous candidate for this category. Since such propositions are purely mathematical, and at least not trivially truth-indeterminate, they are more relevant to the question of the second disjunct of Gödel's disjunction as understood by Gödel than is the Knower sentence. However, there are at present no arguably absolutely undecidable mathematical sentences that are *uncontroversially* truth-determinate. Indeed, the question of whether there are truth-indeterminate set-theoretic propositions turns on deep and unresolved foundational questions. In any case, as mentioned earlier, set-theoretic statements fall outside the scope of the present chapter.

11.3 A New Disjunction

In this section, we introduce a variant of Gödel's disjunction in the language of *EA* that will be investigated in the remainder of the chapter. In order to do so, §11.3.1 will present the principle that has been called in the literature *Epistemic Church's Thesis*, and §11.3.2 will explore the connection between such a principle and Gödel's disjunction. Finally, §11.4 will test *ECT* in a class of models which embed certain idealisations.

11.3.1 Epistemic Church's Thesis

Since functions are infinite abstract objects, human subjects—even in the idealised sense—do not have epistemic access to functions independently of the interpreted linguistic expressions that denote them (call these *function presentations*). According to a thesis proposed in [Shapiro 1985b], a function presentation F is *calculable* if and only if there is an algorithm A such that it is *a priori* knowable that A represents F, where calculability is a "pragmatic counterpart" of the notion of computability and directly involves the human ability of computing a function [Shapiro 1985b, pp. 42–43].

This leads Shapiro to define *calculability* as follows:

Thesis 3 *A function f is calculable if and only if, recognisably, for every number m given in canonical notation, a canonically given number n exists such that the statement $f(m) = n$ is absolutely provable.*

Using Shapiro's notion of calculability, we can express in $\mathcal{L}_{EA}$ a variant of the Church–Turing thesis [Shapiro 1985b, p. 31]:

Thesis 4 (*ECT*)

$$\Box\forall x\exists y\Box\phi(x,y) \rightarrow \exists e\,[e \textit{ is a Turing machine } \wedge \forall x : \phi(x, e(x))],^{7}$$

for ϕ ranging over formulae of the language of epistemic arithmetic.[8]

This principle is called *Epistemic Church's Thesis* (*ECT*) in the literature. Note that in order for the antecedent to ensure that $\phi(x,y)$ expresses a *function*, a choice principle is implicit in *ECT*. However, the choice principle can be eliminated by prefixing the functionality of $\phi(x,y)$ as a condition on *ECT*, so that it assumes the form

$$\phi(x,y) \text{ is functional } \rightarrow \textit{ECT}.$$

Shapiro considers *ECT* "a weaker version of *CT* [in the standard formalisation] which is closer to Church's thesis [than the intuitionistic version of the latter]" [Shapiro 1985b, p. 31].[9] That *ECT* is closer to the Church–Turing thesis than the intuitionistic variant of Church's thesis is due to the fact that, as in the Church–Turing thesis, the existential quantifier in the consequent of *ECT* is *classical*. Thus, *ECT* does not require that any specific Turing machine can be *shown* to compute the effectively computable function described in the antecedent.

Nonetheless, there are strong reasons to be sceptical that *ECT* approximates the content of the Church–Turing thesis in *EA*. The antecedent of *ECT* does not involve the informal notion of algorithm, so it is implausible to consider the antecedent of *ECT* as expressing that $\phi(x,y)$ is algorithmically computable. Indeed, there is no way to *directly* express or quantify over algorithms in the language of *EA* [Shapiro 1985b, pp. 41–43]. Another reason why *ECT* does not capture the content of the Church–Turing thesis is that the *converse* of the Church–Turing thesis is obviously true, whereas the converse of *ECT* is not obviously true [Black 2000, § 2].[10]

11.3.2 ECT and Absolute Undecidability

The truth of *ECT* (or even *ECT* restricted to *arithmetical* relations) would have a significant consequence, for it entails that there are absolutely undecidable propositions in the language of *EA*:

Proposition 3 *If ECT is true, then there are absolutely undecidable propositions expressible in* $\mathcal{L}_{EA}$.

Proof Suppose that there were no such absolutely undecidable propositions. Then for such sentences, a priori demonstrability would coincide with truth. All such occurrences of *ECT* could therefore be erased without changing the truth values of these sentences. But for any non-computable functional predicate $\phi(x,y)$, the corresponding instance of *ECT* would be false. So, contrapositively, if *ECT* is true, then there must be absolutely undecidable propositions. ■

In other words, we have arrived at a new disjunctive thesis (henceforth, *ND*) that is somewhat reminiscent of Gödel's disjunction:

Thesis 5 (*ND*) *Either ECT fails, or there are absolutely undecidable statements (expressible in* $\mathcal{L}_{EA}$*), or both.*

The antecedent of *ECT* does not express the notion that the human mathematical mind is, or is not, a Turing machine. Therefore, the content of the new disjunctive thesis *ND* is not the same as that of Gödel's disjunction. In fact, it is consistent for *ECT* to hold and for the following to fail:

$$\exists e \forall x \in \mathcal{L}_{EA} : T(\ulcorner \Box x \urcorner) \leftrightarrow \text{“}x \text{ is enumerated by the Turing machine } e\text{”},$$

where T is a truth predicate for $\mathcal{L}_{EA}$.

Proposition 3 can be sharpened by weakening its assumption and strengthening its conclusion:

Theorem 1 *If ECT restricted to* Π_1 *arithmetical relations* $\phi(x,y)$ *holds, then there are absolutely undecidable* Π_3 *sentences of* $\mathcal{L}_{EA}$.

Proof We prove the contrapositive, that is: If there are no Π_3 absolutely undecidable sentences of $\mathcal{L}_{EA}$, then *ECT* restricted to Π_1 arithmetical relations is false.

Suppose that there are no absolutely undecidable Π_3 sentences in $\mathcal{L}_{EA}$:

$$\Box\Psi \leftrightarrow \Psi \text{ for all } \Pi_3 \text{ sentences } \Psi \in \mathcal{L}_{EA}.$$

Choose a Turing-uncomputable total functional Π_1 arithmetical relation $\phi(x,y)$; from elementary recursion theory we know that such $\phi(x,y)$ exist.

Then, $\forall x \exists y \phi(x,y)$. But then we also have that $\forall x \exists y \Box \phi(x,y)$. The reason is that $\Pi_1 \subseteq \Pi_3$, so for every m and n, $\phi(m,n)$, being a Π_1 statement, entails $\Box\phi(m,n)$. However, $\forall x \exists y \Box \phi(x,y)$ is now a Π_3 statement of $\mathcal{L}_{EA}$, so again from our assumption it follows that $\Box \forall x \exists y \Box \phi(x,y)$.

Therefore, for the chosen $\phi(x,y)$, the antecedent of *ECT* is true whereas its consequent is false. Therefore, for the chosen $\phi(x,y)$, *ECT* is false. ■

From the existence of absolutely undecidable propositions, certain consequences can be derived.

Recall McKinsey's axiom (the **S4.1** axiom) for propositional modal logic, which states (roughly) that there are no *provably* absolutely undecidable statements:

Axiom 2 (McKinsey) $\neg\Box(\neg\Box\phi \wedge \neg\Box\neg\phi)$.

Corollary 1 *From the absolute provability of ECT restricted to* Π_1 *relations, it follows that McKinsey's axiom for absolute provability fails.*

Proof The proof of Theorem 3 produces a concrete absolutely undecidable sentence from instances of *ECT*. If these instances can be "necessitated," then the proof yields a provably undecidable sentence, which contradicts McKinsey's axiom. ■

In the antecedent of this corollary, *ECT* is taken as an axiom scheme, the instances of which fall within the scope of the necessitation axiom.[11] So the corollary says that *if ECT* can be established by *a priori* means, then even though we probably cannot establish that there are arithmetical absolutely undecidable sentences (Section 11.2.2.2), we will have a proof of the existence of absolutely undecidable sentences in the language of *EA*. Or, in other words, in this situation, the **S4.1** axiom may well hold for arithmetical sentences, but not for all sentences of $\mathcal{L}_{EA}$.

Corollary 2 *If ECT holds, then the converse of ECT fails.*

Proof Suppose that *ECT* holds; then there are absolutely undecidable propositions. Let ψ be an absolutely undecidable proposition. Define $\phi(x,y)$ as

$$[(x = x \wedge y = 1) \leftrightarrow \psi] \wedge [(x = x \wedge y = 0) \leftrightarrow \neg\psi].$$

Then $\phi(x,y)$ defines either the constant 1 function or the constant 0 function, whereby $\exists e \forall x \phi(x, e(x))$. But since ψ is absolutely undecidable, we have that

$$\neg\Box\forall x\exists y\Box\phi(x,y).$$ ■

Earlier we observed that in contrast to the converse of the Church–Turing thesis, the converse of *ECT* is not obviously true. Now we see that the situation is in fact worse: if *ECT* is true, then its converse is plainly false. This brings in sharper relief the fact that *ECT* does not have the same content as the Church–Turing thesis.

Note that the proof of Corollary 2 also entails that there are coextensive functional relations $\phi(x,y)$ and $\psi(x,y)$ expressible in $\mathcal{L}_{EA}$ such that $\Box\forall x\exists y\Box\phi(x,y)$ is true, whereas $\Box\forall x\exists y\Box\psi(x,y)$ is false. That is, it also follows as a corollary that Shapiro's notion of calculability is an intensional notion:

Corollary 3 *If ECT holds, then the antecedent of ECT is intensional.*

Proof Suppose that *ECT* holds; then there are absolutely undecidable propositions. Let ψ be an absolutely undecidable true proposition. Let g be defined and calculable. Then we can define a function f in terms of g and ψ:

$$\begin{cases} \forall x f(x) := g(x) \text{ if } \psi \\ \forall x f(x) := g(x) + 1 \text{ if } \neg\psi. \end{cases}$$

Then we have that f is co-extensional with g:

$$\forall x f(x) = g(x)$$

because ψ is true. Thus, f is *not* calculable because in order to compute f we have to know whether ψ is true. ■

Notice that f is not *provably* coextensive with g, because if we could prove that $\forall x : f(x) = g(x)$, then we would be able to prove that ψ is true.

11.4 Models for *ECT*

In this section, we will discuss the status of *ECT*. In §11.4.1, we will briefly survey some mathematical facts about *ECT* and we will discuss whether it is possible to have a conclusive *a priori* argument for the truth of *ECT*. In §11.4.2 and §11.4.3 we will evaluate *ECT* in models for $\mathcal{L}_{EA}$.

11.4.1 Is ECT True?

Some key metamathematical facts are known about *ECT*. The principle *ECT* was shown to be consistent with *EA* in [Flagg 1985], so it must have models. The proof of the consistency of *ECT* with *EA* was simplified in [Goodman 1986], and then further in [Horsten 1997]. In [Halbach and Horsten 2000] it was shown that from instances of *ECT* as *hypothesis*, no arithmetical theorems can be proved that are not already theorems of Peano arithmetic. As far as we are aware, the question of whether when *ECT* is added as an *axiom* to *EA* this yields an arithmetically conservative extension of Peano arithmetic is still open. It is also known that *EA* + *ECT* has the disjunction property and the numerical existence property [Halbach and Horsten 2000].

The disjunctive principle *ND* that was argued for in §11.3.2 bears witness to the fact that the truth of *ECT* has significant philosophical consequences. *If ECT* is true, then there are non-contingent absolutely undecidable propositions that are perfectly truth-determinate. However, the most important question that is still open is whether *ECT is* a true principle (despite the fact that it is *not* a faithful formalisation of the Church–Turing thesis).

In the discussion about the Church–Turing thesis, a distinction is often made between *quasi-empirical* (*a posteriori*) and conceptual (*a priori*) evidence for the thesis. For a long time it was thought that the Church–Turing thesis cannot be proved, and that all our evidence for its truth is quasi-empirical. In particular, it has been argued that all reasonable

attempts that have been proposed to capture the notion of algorithmically computable function have turned out to be extensionally provably equivalent, and that therefore we have strong inductive grounds for thinking that we have mathematically captured the notion of algorithm.

In recent decades, however, most scholars have come to believe that it is rather Turing's [1936] conceptual analysis of the notion of algorithm that gives us conclusive *a priori* evidence for the Church–Turing thesis.[12] Similar questions to those about evidence for the Church–Turing thesis can be asked about our evidence for *ECT*. More specifically, one can ask whether a convincing *a priori* argument for *ECT* can be found, or whether only weaker (and perhaps quasi-empirical, *a posteriori*) forms of evidence can be found for it, or whether we could construct putative counterexamples to *ECT*.

One might try to use the Church–Turing thesis to argue that *ECT* is true. In [Horsten 1998, §4, p. 15] the following argument was formulated. Suppose that the following thesis holds:

Thesis 6 *The only way in which a statement of the form* $\forall x \exists y \Box \phi(x,y)$ *can be proved is by giving an algorithm for computing* ϕ.

It follows from the Church–Turing thesis that the function expressed by $\phi(x,y)$ is Turing-computable. Given Turing's *a priori* evidence for the Church–Turing thesis, it then follows that if we have good *a priori* evidence for Thesis 6, we will have an *a priori* argument for *ECT*.

Unfortunately, evidence for Thesis 6 is lacking. We have at present no way of excluding that for some functional relation $\phi(x,y)$, it is absolutely provable *in a non-constructive way* that $\forall x \exists y \Box \phi(x,y)$. Such a proof would not involve an algorithm for generating, for every x, a proof of $\phi(x,y)$ for some y. In sum, the argument of [Horsten 1998, §4] does not carry conviction,[13] and thus the prospects for having strong *a priori* arguments for ECT are not promising.

11.4.2 *Simple Machines*

We can try to test *ECT* in some models that incorporate reasonable-looking idealisations on the notion of *a priori* knowability: it is possible that *ECT* will hold (or will fail to hold) in a variety of models that embody reasonable idealisations,[14] and if this is the case, then we would have some evidence in support of (or against) *ECT*. To this task we now turn.

Theories formulated in $\mathcal{L}_{EA}$ can themselves be regarded as models. Given a theory S, we define *truth in* S ($S \models \ldots$) as follows:

- The interpretation of the arithmetical vocabulary is standard.
- The interpretation of the classical Boolean connectives is as usual.
- $S \models \Box\phi \Leftrightarrow \phi \in S$.[15]

This will be called the *theory-as-model* perspective. We have argued that *EA* is sound for *a priori* knowability or absolute provability; on this grounds, it is reasonable to assume of

theories-as-models S that $EA \subseteq S$. More generally, we will be interested in theories that are *sound* in the following precise sense:

Definition 1 *A theory S is called* sound *if* $S \models S$.

Using Shapiro's adaptation of the Kleene slash [Shapiro 1985b, p. 18], which we shall not rehearse here, it can easily be seen that *EA* is a sound theory:[16]

Theorem 2 $EA \models EA$.

For obvious reasons, we will say that *EA* is the *minimal* model for *EA*.

EA also makes *ECT* true. In order to show this, we first recall that *EA* has the epistemic analogue of the *numerical existence property* (henceforth, *NEP*) [Shapiro 1985a, pp. 19–20]:

Theorem 3 (NEP) $EA \vdash \exists x \Box \phi(x) \Rightarrow$ *there is an* $n \in \mathbb{N}$ *such that* $EA \vdash \phi(\overline{n})$,

with $\phi(x)$ being a formula with only x free. It should be noted that the antecedent does not simply express that it is *a priori* knowable that the extension of ϕ is not empty ($\Box(\exists x \phi(x))$), but it expresses the stronger statement that there is a *particular* number x such that it is knowable that x has the property ϕ. So the theorem tells us that if *EA* proves the existence of a number x of which it is *a priori* knowable that it has the property ϕ, then there is a particular number x such that *EA* proves that it is *a priori* knowable that x satisfies the property ϕ.

We can now show that *EA* makes *ECT* true:

Theorem 4 $EA \models ECT$.

Proof $EA \models \Box \forall x \exists y \Box \phi(x,y) \Rightarrow (T)$
$EA \vdash \forall x \exists y \Box \phi(x,y) \Rightarrow (NEP)$
$\forall m \exists n : EA \vdash \phi(\overline{m}, \overline{n})$.

Now let e be the Turing machine that successively for each m finds the shortest *EA*-proof of $\phi(\overline{m}, \overline{n})$ for some minimal n, and outputs n. Then by the soundness of *EA* we have that

$$EA \models \exists e \forall x \exists z \exists v : T(e,x,z) \wedge U(z,v) \wedge \phi(x,v),$$

where T is Kleene's T-predicate and U is Kleene's U-function. ∎

So *EA*, seen as a model, is also the simplest model for *ECT*. In [Halbach and Horsten 2000] it is shown that $EA + ECT \models NEP$. Generalising from this, we can see the following:

Theorem 5 *For all r.e. sound* $S \supseteq EA$ *that have NEP,*

$$S \models ECT.$$

Hence, there are many simple models of *ECT* that assign a recursively enumerable extension to $\Box$. Note that because of the independence of *ECT* from *EA* (see [Flagg 1985] and [Carlson 2015]), we have that

$$EA \not\models \Box ECT.$$

Indeed, it is known that models that also make $\Box ECT$ true are necessarily somewhat complicated [Carlson 2015]. However, this need not unduly concern us here, because, as we have seen, it is anyway somewhat difficult to see how we can *a priori* come to know *ECT*.

11.4.3 *More Realistic Models?*

In this section, we will construct somewhat more *realistic* models for the behaviour of an idealised mathematical community. The aim is to construct simple models for *ECT* that do not necessarily assign a recursively enumerable extension to the $\Box$ operator in order to allow for the possible non-systematicity of the cumulation of knowledge over time, and to test whether *ECT* is true in a wide class of such models.

We start by defining possible worlds models for $\mathcal{L}_{EA}$. We base our models on a *branching time* framework. The informal idea behind this is as follows. A possible world, or possible space-time, might be seen as a linear sequence of moments at which new proofs are generated. These linear structures may be taken to partially overlap in such a way that the union of all the possible space-time moments form a tree (partial ordering) under the earlier-than relation.

Definition 2 *A frame $\mathcal{F} = \langle \mathcal{T}, < \rangle$ is a partial ordering relation.*

The elements of $\mathcal{T}$ can be seen as possible moments in idealised branching time. Instead of writing $t_i < t_j$ for $t_i, t_j \in \mathcal{T}$, we will often simply write $i < j$.

Definition 3 *A possible worlds model $\mathcal{M}$ is a structure of the form $\langle \mathcal{F}, f \rangle$, with $\mathcal{F}$ a frame and $f : \mathcal{T} \mapsto \mathcal{P}(\mathcal{L}_{EA})$.*

Informally, $f(t_i)$ specifies the theory that is known at moment t_i.

Even though we have defined truth in a theory-as-model, we have not yet defined truth in a possible worlds model. Let us do that now:

- The interpretation of the arithmetical vocabulary is standard.
- The interpretation of the classical Boolean connectives is as usual.
- $\mathcal{M} \models \Box\phi \Leftrightarrow \phi \in \bigcup_i f(t_i)$.

Then we immediately see:

Proposition 4 $\mathcal{M} \models \Box\phi \Leftrightarrow \bigcup_i f(t_i) \models \Box\phi$.

This proposition connects the possible worlds perspective with the theory-as-model perspective.

To ensure that the appropriate degree of idealisation is satisfied, we require that possible worlds models meet the following conditions for all i, j:

I. (**Closure**) $f(t_i)$ is closed under logic and contains the theorems of *EA*.

II. (**Cumulativity**) $\phi \in f(t_i) \Rightarrow \forall k > i : \phi, \psi \in f(t_k)$.

III. (**Positive Introspection**) $\phi \in f(t_i) \Rightarrow \forall j > i : \Box\phi \in f(t_j)$.

IV. (**Soundness**) $\phi \in f(t_i) \Rightarrow f(t_i) \models \phi$ and $\bigcup_i f(t_i) \models \phi$.

V. (**Finiteness**) $f(t_i)$ is recursively enumerable.

The requirement of soundness (IV) entails that each $f(t_i)$ has the numerical existence property (*NEP*).

The requirements of closure (I) and cumulativity (II) are idealisations. It is assumed that the mathematical community deduces the logical consequences of what it knows, and it is assumed that the mathematical community has a perfect memory. The requirement of positive introspection (III) is a reflective property. If a subject knows ϕ *a priori*, then they can, in the following moment in time, reflect on their grounds for believing ϕ *a priori* and conclude (*a priori*) that they are strong enough to warrant *a priori* knowledge that ϕ. In other words, they know *a priori* that they know ϕ *a priori*. The soundness requirement (IV) might be argued for from the definition of the concept of knowability: a subject can only come to know *a priori* at a certain moment in time that ϕ if ϕ is true, for it is analytic of the concept of knowledge that it entails the truth of what is known. Condition (V) is a finiteness requirement since any r.e. theory is finitely axiomatisable in a language extension. It is motivated by the fact that since the human mind (even the mind of an idealised mathematical community) is finite at every given moment in time, the content of what is *a priori* implicitly known (given closure under logical consequence) is contained in a Turing machine.

We now see that for such $\mathcal{M}$ the following holds:

Theorem 6

1. $\mathcal{M} \models$ *EA except for the K-axiom.*
2. $\mathcal{M} \models$ *ECT.*

Proof

(1.) In order to show that $\mathcal{M} \models EA$, we show that $\mathcal{M} \models 4, T$. (We already know that $\mathcal{M}$ models the necessitations of these principles, by the closure condition on the $f(t_i)$'s.) That $\mathcal{M} \models 4$ follows from positive introspection; the fact that $\mathcal{M} \models T$ follows from the soundness condition.

(2.) $\mathcal{M} \models \Box\forall x\exists y\Box\phi(x,y) \Rightarrow (T)$
$\exists i : f(t_i) \vdash \forall x\exists y\Box\phi(x,y) \Rightarrow (NEP)$
$\exists i\forall m\exists n : f(t_i) \vdash \phi(\overline{m},\overline{n})$.

Now let e be the Turing machine that successively for each m finds the shortest $f(t_i)$-proof of $\phi(\overline{m},\overline{n})$ for some minimal n, and outputs n. Then by the soundness condition (see IV above) on $\mathcal{M}$ we have that

$$\mathcal{M} \models \exists e\forall x\exists z\exists v : T(e,x,z) \wedge U(z,v) \wedge \phi(x,v).$$ ■

Note that if the frame of $\mathcal{M}$ is a total ordering, then also the K-axiom holds (by cumulativity), and then $\mathcal{M} \models EA$.

Theorem 6 shows that there are many possible worlds models of *EA* that make *ECT* true. A particularly simple and intuitive subclass of possible worlds models consists of those in which the frame is an ω-sequence; let us call these models *ω-sequence models.*[17] This may be taken to be a particularly natural idealised scenario insofar as it depicts discrete linear time going on indefinitely, which seems the appropriate idealisation of the actual structure of time.

Even though it is assumed that each $f(t_i)$ is recursively enumerable, it is *not* assumed that the extension of $\Box$ in the model as a whole (i.e. $\bigcup_i f(t_i)$) is recursively enumerable: there are many intuitive branching time models according to which there is a non-recursively enumerable collection of *a priori* knowable sentences of $\mathcal{L}_{EA}$ in which *ECT* is nonetheless true. Indeed, requirement (V) is no restriction at all on the complexity of the content of what is *a priori* knowable. Let $\langle \phi_i \rangle_i$ be an enumeration of the set of true sentences of the language of arithmetic; then, if we let the extension of *a priori* knowledge at time i be the logical closure of $\{\phi_1, \ldots, \phi_i\}$, the constraint is satisfied, while this entails that the extension of *a priori* knowability over time is the collection of all arithmetical truths.

The possible worlds models that we have discussed above seem to incorporate reasonable idealisations on *a priori* knowability. There is, however, a worry about the justification of one part of the soundness requirement (IV). The statement $\phi \in f(t_i) \Rightarrow \bigcup_i f(t_i) \models \phi$ is indeed supported by the fact that a subject can only come to know *a priori* at a certain moment in time that ϕ, if ϕ is true *from a timeless perspective.* However, it is unclear how we can justify that $\phi \in f(t_i) \Rightarrow f(t_i) \models \phi$. The concern here goes back to the worries that we expressed about Thesis 6 in §11.4.1. Might it not be the case that at some stage the mathematical community has proved in a *non-constructive* manner that $\exists x \Box \phi(x)$, that is, that it has proved this existential statement without producing a witness? However, we have seen that the doubtful part of the soundness requirement is needed to ensure that each $f(i)$ has the numerical existence property, which is in turn used in the proof of Theorem 6. Thus, the class of models which we have proved to validate *ECT* is perhaps not as broad as one might wish.

A specific class of models that one can consider is the class of models in which the extension of absolute provability is given by a *systematic* transfinite progression of formal theories in the sense of [Feferman 1962]. Here subsequent systems are generated from earlier ones by adding uniform reflection principles to what has already been obtained. These systems were the focus of [Kreisel 1972], in which Kreisel tried to determine the truth value of principles in the vicinity of (but probably not quite identical to) *ECT*. We have reserved a detailed discussion of these models for another occasion; let it suffice to say here that in all such models, *ECT* holds (as does *EA*).[18]

11.5 Conclusion

Gödel's disjunction is generally taken to have been shown to be true. But until now we have no compelling evidence for or against any of its two disjuncts.

In this chapter, we have investigated a related disjunctive thesis according to which either Epistemic Church's Thesis (*ECT*) is false, or there are absolutely undecidable propositions expressible in the language of epistemic arithmetic.

It has emerged that this new disjunctive thesis is in the same boat as Gödel's disjunction. The new disjunction is unassailable, but we have no convincing philosophical arguments for or against each of its disjuncts taken individually. In particular, at present we have no convincing *a priori* argument for *ECT*. And in the absence of such an argument, it also seems difficult to find an *a priori* argument for the thesis that there are absolutely undecidable propositions. So it is not immediately obvious how the new disjunctive thesis can be seen as a stepping stone to *a priori* knowledge about the limits of the extension of the notion of *a priori* knowability.

We therefore went on to "test" *ECT* in models for the language of epistemic arithmetic. It turns out that in a wide class of such models, *ECT* holds. However, the significance of this finding is limited by the fact that it is built into these models that there are no non-constructive ways of proving statements of the form $\forall x \exists y \Box \phi(x,y)$; we have seen that it is difficult to argue convincingly for (or against) this assumption. In the end, we must therefore conclude that it would be premature to claim that there is quasi-empirical evidence for the truth (or for the falsity) of *ECT*. At present we therefore also lack evidence for the existence of absolutely undecidable propositions expressible in the language of epistemic arithmetic.

Acknowledgments

Earlier versions of this article were presented at the *Plurals, Predicates and Paradoxes* seminar at Birkbeck College, University of London, at the *Logic Colloquium* in Manchester, at the *Midwest Philosophy of Mathematics Workshop* at the University of Notre Dame, at the *Philosophy of Mathematics Seminar* at the University of Oxford, and at the *Intensionality in Mathematics* conference at the University of Lund: thanks to the audiences for valuable reactions. We are grateful to Roy Cook, Jan Heylen, Harold Hodes, Peter Koellner, Richard Pettigrew, Philip Welch, and Tim Williamson for insightful comments.

Notes

1. For an extended discussion of the implications and non-implications of Gödel's theorems, see [Franzen 2005].
2. For a discussion of the different viewpoints, see [Williamson 2000, chapter 12].
3. See [Antonutti Marfori and Horsten subm] and [Kreisel 1972].
4. For a discussion, see [Horsten 1997, p. 640].
5. Of course, a provability predicate with the properties assumed in Proposition 2 cannot be *proven* to exist in *EA*.
6. See e.g. the debate in *The Bulletin of Symbolic Logic* **6**(4), 2000.
7. The notion of being a Turing machine can be formalised in the background language of arithmetic in the standard way in terms of Kleene's T-predicate and the U function symbol.
8. This is not necessary, though; it suffices for the purposes of this paper that functional predicates range over formulas in the language of arithmetic, or even over a fragment of this language.
9. For a discussion of the intuitionistic version of Church's thesis, see [Troelstra and van Dalen 1988].

10. Note that the converse of the Church–Turing thesis entails that if *ECT* holds, then for every functional $\phi \in \mathcal{L}_{EA}$: if ϕ is calculable, then the graph of ϕ is algorithmically computable. (Thanks to Jan Heylen for pointing this out.)
11. If *ECT* is instead treated as a *hypothesis*, then the conclusion of this corollary does not follow. For a more extensive discussion of the treatment of *ECT* as an axiom scheme versus its treatment as a hypothesis, see [Halbach and Horsten 2000] and [Horsten 2006].
12. For an extended discussion, see [Sieg 1994] and [Soare 1996]. Actually, it was Gandy who first (in print) argued that Turing had given an *a priori* proof of the Church–Turing thesis: see [Gandy 1988].
13. For a discussion of this argument, see [Horsten 2006].
14. See [Antonutti Marfori 2010] and [Antonutti Marfori 2013] for discussions of what might be reasonable idealisations built into the notion of provability in principle. For a contrasting view, see [Williamson 2015] in this volume.
15. Since we are quantifying into the context of $\Box$, what will be said below must be relativised to assignments: see [Alexander 2013] and [Heylen 2013]. For ease of reading, we will ignore this complication in what follows.
16. Carlson denotes this by saying that *EA* is a *machine* [Carlson 2000].
17. In ω-sequence models, the cumulativity condition follows from positive introspection and closure.
18. See [Antonutti Marfori and Horsten subm].

References

[Alexander 2013] Alexander, S.A. A machine which knows its own code. *Studia Logica* **102**, pp. 567–576, 2014.

[Anderson 1983] Anderson, C.A. The paradox of the knower. *Journal of Philosophy* **80**, pp. 338–355, 1983.

[Antonutti Marfori 2010] Antonutti Marfori, M. Informal provability and mathematical rigour. *Studia Logica* **96**, pp. 261–272, 2010.

[Antonutti Marfori 2013] Antonutti Marfori, M. *Theories of Absolute Provability*. Ph.D. Thesis, University of Bristol, 2013.

[Antonutti Marfori and Horsten subm] Antonutti Marfori, M. and Horsten, L. Human effective computability. Submitted, 2015.

[Black 2000] Black, R. Proving Church's Thesis. *Philosophia Mathematica* **8**, pp. 244–258, 2000.

[Boolos 1982] Boolos, G. Extremely undecidable sentences. *Journal of Symbolic Logic* **47**, pp. 191–196, 1982.

[Carlson 2000] Carlson, T. Knowledge, machines, and the consistency of Reinhardt's strong mechanistic thesis. *Annals of Pure and Applied Logic* **105**: pp. 51–82, 2000.

[Carlson 2015] Carlson, T. Can a machine know that it is a machine? This volume.

[Enderton 2001] Enderton, H.B. *A Mathematical Introduction to Logic*, Second Edition. San Diego, California, Academic Press, 2001.

[Feferman 1962] Feferman, S. Transfinite recursive progressions of formal theories. *Journal of Symbolic Logic* **27**, pp. 259–316, 1962.

[Feferman and Solovay 1990] Feferman, S. and Solovay, R. Introductory note to 1972a. In: S. Feferman et al. (eds.) *Kurt Gödel. Collected Works. Volume II: Publications 1938–1974*, pp. 281–304, Oxford: Oxford University Press, 1990.

[Fitch 1963] Fitch, F. A logical analysis of some value concepts. *Journal of Symbolic Logic* **28**, pp. 135–142, 1963.

[Flagg 1985] Flagg, R. Church's Thesis is consistent with Epistemic Arithmetic. In [Shapiro 1985a, pp. 121–172].

[Folina 1998] Folina, J. Church's Thesis: prelude to a proof. *Philosophia Mathematica* **6**, pp. 302–323, 1998.

[Franzen 2005] Franzen, T. *Gödel's Theorem: An Incomplete Guide to its Use and Abuse*. A.K. Peters, Wellesley, MA, 2005.

[Gandy 1988] Gandy, R. The confluence of ideas in 1936. In: R. Herken (ed.) *The Universal Turing Machine: A Half Century Survey*, pp. 55–111, New York: Oxford University Press, 1988.

[Gödel 1951] Gödel, K. Some basic theorems on the foundations of mathematics and their implications. [1951] In: S. Feferman et al. (eds.) *Kurt Gödel. Collected Works. Volume III: Unpublished Essays and Lectures*, pp. 304–323, Oxford: Oxford University Press, 1995.

[Goodman 1986] Goodman, N., Flagg realizability in Epistemic Arithmetic. *Journal of Symbolic Logic* **51**, pp. 387–392, 1986.

[Halbach and Horsten 2000] Halbach, V. and Horsten, L. Two proof-theoretic remarks about *EA* + *ECT*. *Mathematical Logic Quarterly* **46**, pp. 461–465, 2000.

[Heylen 2013] Heylen, J. Modal-epistemic arithmetic and the problem of quantifying in. *Synthese* **190**, pp. 89–111, 2013.

[Horsten 1997] Horsten, L., Provability in principle and controversial constructivistic principles. *Journal of Philosophical Logic* **26**, pp. 635–660, 1997.

[Horsten 1998] Horsten, L. In defense of Epistemic Arithmetic. *Synthese* **116**, pp. 1–25, 1998.

[Horsten 2006] Horsten, L. Formalizing Church's Thesis. In: A. Olszewski et al. (eds.) *Church's Thesis after 70 years*. Heusenstamm: Ontos Verlag, 2006.

[Kreisel 1972] Kreisel, G. Which number theoretic problems can be solved in recursive progressions on Π^1_1-paths through O? *Journal of Symbolic Logic* **37**, pp. 311–334, 1972.

[Lucas 1961] Lucas, J.R. Minds, machines and Gödel. *Philosophy* **96**, pp. 112–127, 1961.

[Myhill 1960] Myhill, J. Some remarks on the notion of proof. *Journal of Philosophy* **57**, pp. 461–471, 1960.

[Penrose 1989] Penrose, R. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press, 1989.

[Penrose 1994] Penrose, R. *Shadows of the Mind. A Search for the Missing Science of Consciousness.* Oxford: Oxford University Press, 1994.

[Reinhardt 1986] Reinhardt, W. Epistemic theories and the interpretation of Gödel's incompleteness theorems. *Journal of Philosophical Logic* **15**, pp. 427–474, 1986.

[Shapiro 1985a] Shapiro, S. *Intensional Mathematics.* Amsterdam: North-Holland, 1985.

[Shapiro 1985b] Shapiro, S. Epistemic and intuitionistic arithmetic. In: [Shapiro 1985a, pp. 11–46].

[Sieg 1994] Sieg, W. Mechanical procedures and mathematical experience. In A. George (ed.), *Mathematics and Mind*, New York: Oxford University Press, 1994.

[Soare 1996] Soare, R. I. Computability and recursion. *Bulletin of Symbolic Logic* **2**, pp. 284–321, 1996.

[Troelstra and van Dalen 1988] Troelstra, A. and van Dalen, D. *Constructivism in Mathematics. An Introduction. Volume 1* Amsterdam: North-Holland, 1988.

[Turing 1936] Turing, A.M. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **42**, pp. 230–265, 1936.

[Turing 1939] Turing, A.M. Systems of logic defined by ordinals. *Proceedings of the London Mathematical Society* Ser. 2, **45**: pp. 161–228, 1939.

[Williamson 2000] Williamson, T. *Knowledge and its Limits.* Oxford: Oxford University Press, 2000.

[Williamson 2015] Williamson, T. Absolute provability and safe knowledge of axioms. This volume.