THE JOURNAL OF PHILOSOPHY

VOLUME CXV, NO. 7, JULY 2018

ON THE QUESTION OF WHETHER THE MIND CAN BE MECHANIZED, I: FROM GÖDEL TO PENROSE*

In this paper I want to address the question of whether "the mind can be mechanized." This is a question with a long history. But it is also a question on which it has been difficult to make genuine progress. For without a precise analysis of the concept of "mechanism," it would be hard to even get started, and, even if one did have such an analysis, it would be hard to see how one could give a definitive argument for or against the claim that the mind can be mechanized.

The situation changed somewhat in the 1930s, through two major developments in mathematical logic. The first development was Turing's analysis of the notion of "computability." Turing gave a convincing analysis of the vague and informal notion of "being computable by an idealized finite machine" in terms of the precise mathematical notion of "being computable by a Turing machine." This enabled one to sharpen the question of whether "the mind can be mechanized" by first focusing on the more specific question of whether "the mathematical outputs of the idealized human mind can coincide with the mathematical outputs of an idealized finite machine," and then sharpening the vague notion of "an idealized finite machine" in terms of

*I am grateful to Leon Horsten and Philip Welch for inviting me to speak on this subject at Bristol, first on March 19, 2010, and then as part of their *Workshop on the Scope and Limits of Mathematical Knowledge*, March 30–31, 2013. I benefited from the comments I received at those talks, as well as from the comments I received when I spoke on the subject at Barcelona, Columbia, Duke, MIT, Pittsburgh, Stanford, and Jerusalem. In particular, I am grateful to Yuri Gurevich for drawing the allusion to Kafka, and to Michael Friedman for an illuminating conversation in which he pressed me to say more about what these results might tell us about the nature of reason. Finally, I would like to thank Samuel Alexander, Sol Feferman, Gabriel Goldberg, Kentaro Fujimoto, Wesley Holliday, Leon Horsten, Hannes Leitgeb, Johannes Stern, Panu Raatikainen, and an anonymous referee for helpful comments on an earlier draft of this paper.

0022-362X/18/1507/337–360

the mathematically precise notion of "a Turing machine." The second development was Gödel's discovery of the incompleteness phenomenon. His incompleteness theorems demonstrated that for any sufficiently strong consistent formal system of mathematics there are mathematical truths that cannot be captured by that formal system. Given the correspondence between formal systems and Turing machines, and given that it appears that we, on the outside, can capture these "missing truths," Gödel's discovery raised the prospect that one might actually use the incompleteness theorems to argue that the mathematical outputs of the idealized human mind do indeed outstrip the mathematical outputs of any idealized finite machine.

This, then, is the question that I would like to address in this paper: Do the incompleteness theorems imply that "the mind cannot be mechanized" (understood in the above sense)? But I would like to stress two things. First, I am addressing a specific version of this question. I am not considering the performance of actual human minds, with their limitations and defects; I am considering the *idealized* human mind and looking at what it can do *in principle*. I am not considering a broad array of outputs that the idealized human mind might have; I am only considering mathematical outputs. I am not considering a wide variety of abilities that the idealized human mind might have, such as the ability to be creative, or form normative judgments, or to fall in love, and so on; I am only considering the ability to generate mathematical outputs. In short, I am only dealing with the specific question of whether "the mathematical outputs of the idealized human mind can coincide with the mathematical outputs of an idealized finite machine."[1] Second, in addition to restricting my attention to this specific version of the question, I will not be considering all possible arguments for the claim that "the mind cannot be mechanized." I will only be dealing with approaches based on the incompleteness theorems. To summarize, I will only be addressing *the question of whether the incompleteness theorems imply that "the mathematical outputs of the idealized human mind do not coincide with the mathematical outputs of any idealized finite machine."*

[1] Throughout this paper I will use the expression "the mind can be mechanized" as shorthand for "the mathematical outputs of the idealized human mind coincide with the mathematical outputs of an idealized finite machine." I want to stress that I do not think that the latter provides an adequate analysis of the former, in the sense of telling us what it *means* to say that "the mind can be mechanized." I am simply focusing on this more specific version of the question and using a convenient shorthand. I will also use "mechanism" as a convenient label for the thesis that the mathematical outputs of the idealized human mind can coincide with the mathematical outputs of an idealized finite mind.

The story begins with Gödel. Gödel thought that his incompleteness theorems had bearing on the question of mechanism, but his position was quite subtle. He did *not* argue that his incompleteness theorems implied that "the mind cannot be mechanized"; he argued, rather, that they implied a weaker, disjunctive conclusion, what I shall call 'Gödel's Disjunction'. The disjunction concerns two central philosophical claims. The first is the claim that we have been considering, namely, the claim that "the mind cannot be mechanized" (understood in our specific sense). The second is the claim that "there are mathematical truths that cannot be proved by the idealized human mind" (or, equivalently, that "there are absolutely undecidable statements"). The disjunction states that at least one of these claims must hold—that is, it states that either "the mind cannot be mechanized" or "mathematical truth outstrips the idealized human mind."

Gödel thought that each disjunct had important philosophical consequences, quite different in each case, but each "very decidedly opposed to materialist philosophy."

> Namely, if the first alternative holds, this seems to imply that the working of the human mind cannot be reduced to the working of the brain, which to all appearances is a finite machine with a finite number of parts, namely, the neurons and their connections. So apparently one is driven to take some vitalistic viewpoint. On the other hand, the second alternative, where there exist absolutely undecidable mathematical propositions, seems to disprove the view that mathematics is only our creation; for the creator necessarily knows all the properties of his creatures, because they can't have any others except those he has given them. So this alternative seems to imply that mathematical objects and facts (or at least *something* in them) exist objectively and independently of our mental acts and decisions, that is to say, [it seems to imply] some form or other of Platonism or "realism" as to the mathematical objects.[2]

Now, it is rarely the case in philosophy that claims are actually established beyond a shadow of a doubt, and this is especially true when those claims concern such large matters as the relationship between mechanism, mind, and mathematical truth. But Gödel—who was generally quite cautious in his claims—went so far as to call the disjunction a "mathematically established fact."[3]

[2] Kurt Gödel, "Some Basic Theorems on the Foundations of Mathematics and Their Implications" (1951), reprinted in *Collected Works, Volume III: Unpublished Essays and Lectures*, ed. Solomon Feferman et al. (New York: Oxford University Press, 1995), pp. 304–23, at p. 311.

[3] *Ibid.*, p. 310.

Our first order of business will be to determine whether the disjunction is indeed a "mathematically established fact."[4]

Let us suppose for the moment that the disjunction is true. The question then arises: Which disjunct holds?

Gödel himself was convinced that the first disjunct is true and the second disjunct is false; that is, he was convinced that the mind cannot be mechanized and that human reason is sufficiently powerful to capture all mathematical truths. But although he was convinced of these stronger claims he did not believe that he was in a position to establish either. He did, however, think that *one day* we would be in a position to prove the first disjunct. What was missing, as he saw it, was an adequate resolution of the paradoxes involving self-applicable concepts like the concept of truth. And he thought that "[i]f one could clear up the intensional paradoxes somehow, one would get a clear proof that mind is not machine."[5] However, he did not think that we had yet arrived at an adequate resolution of the paradoxes, and, lacking such a resolution, he felt that the most he could claim to have established was the disjunctive conclusion.

Others, who have discussed these matters since Gödel, have claimed more. They have claimed that the incompleteness theorems imply that the first disjunct holds.

There are really two different generations of arguments for the first disjunct. The first generation began with Nagel and Newman in 1956 and continued with Lucas in a talk of 1959 and a subsequent publication in 1961.[6] Nagel and Newman's argument was criticized by Putnam, while Lucas's argument was much more widely criticized in the literature.[7] The topic was revisited in 1989 by Penrose in his famous

[4] I will not in this paper directly address Gödel's claims concerning the philosophical significance of each disjunct. However, in the final section of the successor to this paper I will say some things that bear on it.

[5] This quotation is from Hao Wang's reconstruction of his conversations with Gödel: Hao Wang, *A Logical Journey: From Gödel to Philosophy* (Cambridge, MA: MIT Press, 1996), p. 187.

[6] James R. Newman and Ernest Nagel, "Gödel's Proof," *Scientific American*, CXCIV, 6 (June 1956): 1668–95; Ernest Nagel and James R. Newman, *Gödel's Proof*, rev. ed (New York: New York University Press, 2001); John Randolph Lucas, "Minds, Machines and Gödel," *Philosophy*, XXXVI, 137 (1961): 112–27.

[7] Hilary Putnam, "Minds and Machines," in Sidney Hook, ed. *Dimensions of Mind: A Symposium* (New York: New York University Press, 1960), pp. 138–64. It is also worth noting that Gödel was quite unhappy with Nagel and Newman's arguments, for reasons that will be apparent from our account of his view below. See Solomon Feferman, "Gödel, Nagel, Minds, and Machines," this JOURNAL, CVI, 4 (April 2009): 201–19. See Paul Benacerraf, "God, the Devil, and Gödel," *The Monist*, LI, 1 (January 1967): 9–32, for an influential criticism of Lucas.

book *The Emperor's New Mind*.[8] This triggered an avalanche in the literature and in the following year there was an open peer commentary of Penrose's book in *Behavioral and Brain Sciences*.[9]

The second generation of arguments—one argument, really—appeared in another book-length account by Penrose in 1994: *Shadows of the Mind: A Search for the Missing Science of Consciousness*.[10] This argument also received a great deal of attention.[11] Penrose has continued to defend his argument—for example, in his address at the Gödel Centenary in 2006 and the subsequent published version of 2011.[12]

Our second order of business will be to assess the cogency of the arguments for the first disjunct. In this paper I will concentrate on the first generation of arguments, and in the successor to this paper I will address the second generation of arguments.

The approach I shall take is somewhat different from the approach that is customary in the literature. One difficulty with the discussion in the literature is that the background assumptions governing the underlying concepts—most notably, the concepts of "an idealized finite machine" (∼"relative provability"), "the idealized human mind"

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

[9] See, in particular, Roger Penrose, Précis of *The Emperor's New Mind: Concerning Computers, Mind, and the Laws of Physics, Behavioral and Brain Sciences*, XIII, 4 (December 1990): 643–54; George Boolos et al., Open Peer Commentary on *The Emperor's New Mind*, *Behavioral and Brain Sciences*, XIII, 4 (December 1990): 655–91; Roger Penrose, "The Nonalgorithmic Mind," *Behavioral and Brain Sciences*, XIII, 4 (December 1990): 692–706; Martin Davis et al., Continuing Commentary on *The Emperor's New Mind*, *Behavioral and Brain Sciences*, XVI, 3 (September 1993): 611–16; and Roger Penrose, "An Emperor Still without Mind," *Behavioral and Brain Sciences*, XVI, 3 (September 1993): 616–22. For a more recent criticism see Haim Gaifman, "What Gödel's Incompleteness Result Does and Does Not Show," this JOURNAL, XCVII, 8 (August 2000): 462–70.

[10] Roger Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness* (New York: Oxford University Press, 1994).

[11] See, for example, David J. Chalmers, "Minds, Machines, and Mathematics: A Review of *Shadows of the Mind* by Roger Penrose," *Journal Psyche*, II (June 1995): 11–20; Solomon Feferman, "Penrose's Gödelian Argument: A Review of *Shadows of the Mind* by Roger Penrose," *Journal Psyche*, II (May 1995): 21–32; Per Lindström, "Penrose's New Argument," *Journal of Philosophical Logic*, XXX, 3 (June 2001): 241–50; Per Lindström, "Remarks on Penrose's 'New Argument'," *Journal of Philosophical Logic*, XXXV, 3 (June 2006): 231–37; Stewart Shapiro "Incompleteness, Mechanism, and Optimism," *Bulletin of Symbolic Logic*, IV, 3 (September 1998): 273–302; and Stewart Shapiro, "Mechanism, Truth, and Penrose's New Argument," *Journal of Philosophical Logic*, XXXII, 1 (February 2003): 19–42.

[12] Roger Penrose, "Gödel, the Mind, and the Laws of Physics," in Matthias Baaz et al., eds., *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth* (New York: Cambridge University Press, 2011), pp. 339–58.

(∼"absolute provability"), and "truth"—are seldom fully articulated and, as a consequence, it is difficult to assess the cogency of the arguments. One of my goals here is to sharpen the debate by making the background assumptions governing the fundamental concepts explicit. Once we do this, we will be able to pull the entire discussion into a framework where we can establish definitive results of the form: "If the principles governing the fundamental concepts are such-and-such, then there is no hope of proving or refuting the first disjunct."[13]

We will see that there is a natural framework, EA_T, governing the concepts of relative provability, absolute provability, and truth, and that in this framework one can give a rigorous proof of Gödel's Disjunction, thereby vindicating, in some sense at least, Gödel's claim that the disjunction is a "mathematically established fact." We will also see that—for reasons Gödel anticipated—the *particular* arguments of Lucas and Penrose are based on an oversight and, as a result, fail. More generally, we shall see that results of Reinhardt and Carlson show that there is *no* argument for the first disjunct in EA_T. Since the axioms of EA_T would seem to encompass all of the assumptions made by proponents of the first disjunct this places a fundamental limitation on their program. I hope that this puts to rest the first generation of arguments for the first disjunct and that all participants in the debate can agree on this.

My strategy is to be as charitable as possible, for the strength of a criticism is proportional to the degree to which it is charitable. But ultimately, for reasons I give at the end of the successor to this paper, I am skeptical of the very terms in which the debate is formulated. Nevertheless, I think that there is something of value in entering the debate. It is like entering Kafka's castle. We begin by accepting an implausible scenario and from there everything proceeds rationally. My hope is that we can enter the castle together and by dint of pure reason find our way back out again.[14]

[13] In pursuing this enterprise I see myself as doing little more than following and underscoring the importance of the beautiful work of William Reinhardt. He was an astonishing thinker. In every encounter with his work one witnesses an adventurous mind, pushing the limits, reaching for the heavens, but always taking stock and returning to the ground and doing the hard work required to ensure that his findings are made rigorous and communicable to everyone.

[14] The mathematical results I shall be discussing in this paper and its successor are proved in a technical companion and the references therein: Peter Koellner, "Gödel's Disjunction," in Leon Horsten and Philip Welch, eds., *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (New York: Oxford University Press, 2016), pp. 148–88.

I. GÖDEL

Let us begin with an informal discussion of the disjunction, taking Gödel as our guide.

I.1. Preliminaries. The disjunction concerns the concepts of "relative provability," "absolute provability," and "truth," and, in a variant formulation, the related concepts of "an idealized finite machine" and "the idealized human mind." We shall formulate our discussion in terms of the first three concepts.

There is no loss of generality in doing this. For it is assumed by all participants in the debate that the terms are understood in a way such that (1) the concept of what is "relatively provable with respect to a given formal system F" is co-extensive with the concept of what is "producible by the idealized finite machine (a Turing machine) M" (where M is the Turing machine corresponding to F) and (2) the concept of what is "absolutely provable" is co-extensive with the concept of what is "producible by an idealized human mind." The first of these assumptions is uncontroversial; in fact, if we take (as we shall) "formal system" to mean "recursive formal system" then the co-extensiveness of the two concepts is a basic theorem in recursion theory. The second assumption, in contrast, does not admit of anything like a proof, dealing as it does with less definite notions. But it is assumed by all participants in the debate that the two expressions are co-extensive. In any case, for the purposes of this discussion the reader should take these expressions as being interchangeable, and so, for example, when we write something concerning what is "absolutely provable" this can be taken to be interchangeable with what is "producible by the idealized human mind." Notice also that in each case we will only be concerned with outputs; that is, we will be concerned with extension, not intension.

We thus have two variant formulations of the various statements we will be dealing with. For example, in one formulation the statement that "the mind can be mechanized" is the statement that "the statements that are absolutely provable coincide with the statements that are provable relative to a recursive formal system"; and in the other, mentalistic formulation (what I shall call "the variant formulation") it is the statement "the outputs of the idealized human mind coincide with the outputs of an idealized finite machine (a Turing machine)." At times I will pass from one formulation to the other. However, it is important to bear in mind that for the purposes of this discussion the underlying notions are to be understood in such a way that the two formulations are extensionally equivalent.

It will be useful to introduce some abbreviations. In this section we will use 'F' for the set of sentences that are provable relative to

a given formal system (which we will also label with 'F'), 'K' for the set of sentences that are absolutely provable, and 'T' for the set of sentences that are true. For readability we will also write '$K(\varphi)$' for '$\varphi \in K$'. We shall say that F is *correct* if $F \subseteq T$. And we shall assume throughout that $K \subseteq T$.

Our goal is to determine what the incompleteness theorems tell us about the relationship between F, K, and T. Gödel makes three main claims concerning this relationship. So let us start there.

I.2. The Incompleteness Theorems. To fix ideas, consider the standard axiomatic system of arithmetic—the axiomatic system of Peano Arithmetic (PA). The language, L_{PA}, of this system contains the usual *logical symbols* (connectives, quantifiers, and equality) and the following *nonlogical symbols*: the constant symbol '0', the unary function symbol 'S', the binary operation symbols '$+$' and '$\times$', and the binary relation symbol '$\leq$'. The axioms of PA contain such statements as '$S(x) \neq 0$' and '$x+S(y) = S(x+y)$' along with the scheme of mathematical induction.

One of the first questions that arises when one considers a formal system F is whether it is *complete*, that is, whether for every statement φ in the language of the system, either F proves φ, or F proves $\neg\varphi$. In certain cases, the answer is "yes"; for example, there is a complete set of axioms for Euclidean Geometry. But remarkably, in the case of arithmetic, the answer is "no," as demonstrated by the *first incompleteness theorem*.

> *Theorem 1 (Gödel).* Assume that PA is consistent. Then there is a sentence φ in L_{PA} such that PA cannot prove φ, and PA cannot prove $\neg\varphi$.[15]

Such a statement φ is said to be *independent* of PA. If we assume (as we shall) that the concept of truth is such that for every statement φ of arithmetic either φ is true or $\neg\varphi$ is true, then we arrive at the conclusion that (assuming that PA is consistent) there are truths of arithmetic that are beyond the reach of relative provability in PA.

The next question that arises is where such limitations "first" occur. To render this question precise we need a stratification of the statements of arithmetic. The most natural way to do this is in terms of quantifier complexity (which boils down to counting the number of alternating quantifiers), and it leads to a hierarchical classification of the statements of arithmetic, ranging from the simplest up

[15] Strictly speaking, this version of the first incompleteness theorem is a strengthening, due to Rosser, of Gödel's original result. Gödel had to assume more than the consistency of PA—he had to assume that PA was Σ^0_1-*sound*, that is, that for every Σ^0_1-sentence φ, if PA proves φ then φ holds. By selecting a different sentence than the one Gödel selected, Rosser was able to weaken the assumption from Σ^0_1-soundness to consistency.

through layers of increasing complexity. Without going into the details, let us just say that the layers of this hierarchy are denoted Σ^0_1, Π^0_1, Σ^0_2, Π^0_2, ..., Σ^0_n, Π^0_n, ... Our question then becomes: Where in this hierarchy do the above limitations first appear?

It is routine to show that for all Σ^0_1-statements φ, if φ is true, then φ is provable in PA. In other words, PA captures Σ^0_1-truth. So there are no limitations at that level. The next possible place at which limitations can appear is at the level Π^0_1 and so the question arises: Does PA capture Π^0_1-truth?

The statement that Gödel produced in (his version of) the first incompleteness theorem is actually a Π^0_1-statement and, upon reflection, one can see that this statement is true. Thus, the answer to our question is: "No, PA does not capture Π^0_1-truth." In fact, the *second incompleteness theorem* provides us with a particularly nice example of a Π^0_1-statement that PA cannot prove, one that is closer to home:

> *Theorem 2 (Gödel).* Assume that PA is consistent. Then PA cannot prove Con(PA).

Here 'Con(PA)' is a statement in L_{PA} which formally renders the informal statement that PA is consistent. The point, for present purposes, is that this is a Π^0_1-statement. Thus, granting the consistency of PA, the second incompleteness theorem shows that PA cannot capture the Π^0_1-truth 'Con(PA)'.[16]

It is important to stress that these theorems are perfectly general. We have stated them for one particular system, namely, PA, but they apply to *any* sufficiently strong formal system; for example, they apply to very weak systems of arithmetic, like Robinson's Q, and they apply to very strong systems of set theory, like ZFC supplemented with large cardinal axioms.

Henceforth we shall restrict our attention to formal systems F which have the minimal amount of strength required for the incompleteness theorems to be in effect.

I.3. The First Claim. Notice that if F is correct then it is consistent, and so by the second incompleteness theorem there is a truth—namely, Con(F)—that is not relatively provable in F. In the words of Gödel, the incompleteness theorems tell us that

[16] If one strengthens the assumption of the theorem to the assumption that PA is Σ^0_1-sound, then one can strengthen the conclusion by adding that PA does not prove $\neg$Con(PA); in other words, under this stronger assumption, Con(PA) is independent of PA.

> no well-defined system of correct axioms [F] can comprise all objective mathematics [T], since the proposition which states the consistency of the system is true, but not demonstrable in the system.[17]

In other words, we have:

> *Claim 1.* For any F,
>
> $F \subseteq T \rightarrow F \subsetneq T$.

The informal reasoning is as follows. If $F \subseteq T$, then $\mathrm{Con}(F) \in T$. But $\mathrm{Con}(F) \notin F$, by the second incompleteness theorem. So $F \subsetneq T$.

This proposition concerns only F and T, each of which is clear and definite. It is thus a clear and definite consequence of the incompleteness theorems.

I.4. The Second Claim. The above conclusion concerns the relationship between F and T. Gödel is careful to note that at this point of his discussion he has said nothing about K. He draws a distinction between "objective mathematics" (by which he means T) and "subjective mathematics" (by which he means K), and he goes on to elaborate the cautionary point about K as follows:

> [O]ne has to be careful in order to understand clearly the meaning of this state of affairs. Does it mean that no well-defined system of correct axioms [F] can contain all of mathematics proper? It does, if by mathematics proper is understood the system of all true mathematical propositions [T]; it does not, however, if one understands by it the system of all demonstrable mathematical propositions [K]. I shall distinguish these two meanings of mathematics as mathematics in the objective [T] and in the subjective [K] sense.[18]

So at this point we have only secured a definite conclusion concerning the relationship between F and T. But our real interest is in the relationship that K bears to F and T. Do the incompleteness theorems tell us anything about this?

Let us say that a statement φ is *relatively undecidable* with respect to F if neither $\varphi \in F$ nor $\neg\varphi \in F$. And let us say that a statement φ is *absolutely undecidable* if neither $\varphi \in K$ nor $\neg\varphi \in K$. In this terminology, Claim 1 tells us that for any sufficiently strong and correct F, the incompleteness theorems provide us with statements that are *relatively* undecidable with respect to F. But are these statements *absolutely* undecidable? Gödel certainly thought that they are *not*:

[17] Gödel, "Some Basic Theorems on the Foundations of Mathematics and Their Implications," *op. cit.*, p. 309.

[18] *Ibid.* Here by "demonstrability" Gödel means "absolute provability."

> [These statements are] not at all absolutely undecidable; rather, one can always pass to "higher" systems in which the sentence in question is decidable. (Some sentences, of course, nevertheless remain undecidable.) In particular, for example, it turns out that analysis is a system higher in this sense than number theory, and the axiom system of set theory is higher still than analysis.[19]

Here is what he had in mind: We know that if PA is consistent then it misses the Π^0_1-truth Con(PA). Let PA_2 be the natural axiomatization of *second*-order arithmetic. It turns out that Con(PA) is provable in PA_2; so, in ascending from PA to PA_2, we capture the Π^0_1-truth that was missed by PA. Of course, the second incompleteness theorem also applies to PA_2, and so, assuming that PA_2 is consistent, it misses the Π^0_1-truth Con(PA_2). But now if we let PA_3 be the natural axiomatization of *third*-order arithmetic we find that it proves Con(PA_2) and so captures the Π^0_1-truth that was missed by PA_2. This pattern continues up through the orders of arithmetic and up through the hierarchy of set-theoretic systems. At each stage a missing Π^0_1-truth is captured and a new one is revealed, and that new Π^0_1-truth is captured at the next stage.

If we grant that each of the systems F in the above hierarchy is subsumed by K—that is, such that $F \subseteq K$—then we can conclude that K outstrips each F in the above hierarchy (for at each successor stage K will capture the Π^0_1-truth missed by the system F at the previous stage). It is tempting to conclude *outright* that K cannot coincide with *any* F, period. But we have to be careful. We have to keep track of our assumptions. Gödel is quite careful—he draws only a *conditional* conclusion:

> For, *it makes it impossible that someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it*: *All of these axioms and rules I perceive* (*with mathematical certitude*) *to be correct, and moreover I believe that they contain all of mathematics*. If someone makes such a statement he contradicts himself. For if he perceives the axioms under consideration to be correct [$K(F \subseteq T)$], he also perceives (with the same certainty) that they are consistent. Hence he has a mathematical insight not derivable from his axioms [$F \subsetneq K$].[20]

In other words, we have:

Claim 2. For any F,

$K(F \subseteq T) \rightarrow F \subsetneq K$.

[19] Gödel, "On Undecidable Sentences," *op. cit.*, p. 35.

[20] Gödel, "Some Basic Theorems on the Foundations of Mathematics and Their Implications," *op. cit.*, p. 309, his italics.

The informal reasoning is as follows: Suppose $K(F \subseteq T)$. We saw above (in the argument for Claim 1) that if $F \subseteq T$ then $\mathrm{Con}(F) \in T$ but $\mathrm{Con}(F) \notin F$. Now, if we also have $K(F \subseteq T)$ then we have $K(\mathrm{Con}(F))$. So $\mathrm{Con}(F) \in K$ but $\mathrm{Con}(F) \notin F$. Thus, $F \subsetneq K$.

I.5. The Third Claim. But can we draw the stronger, *non*-conditional conclusion; that is, can we drop the condition that $K(F \subseteq T)$ and simply conclude outright that K cannot coincide with any F? Gödel is quick to point out that we cannot (at this point at least) draw such a conclusion:

> However, as to subjective mathematics [K], it is not precluded that there should exist a finite rule [F^*] producing all its evident axioms [$F^* = K$]. However, if such a rule exists, we with our human understanding could certainly never know it to be such, that is, we could never know with mathematical certainty that all propositions it produces are correct [that is, we can't have $K(F^* \subseteq T)$].[21]

In other words, (for all we have shown) it may indeed be the case that there is a "master system" F^* such that $F^* = K$. We have only shown that if there is such an F^* then it must be "hidden" in the sense that we cannot absolutely prove that it is correct.[22]

Now, if there *was* in fact such an F^* it would have important implications:

> If it were so, this would mean that the human mind (in the realm of pure mathematics) [K] *is* equivalent to a finite machine [F^*] that, however, is unable to completely understand its own functioning. This inability [of man] to understand himself would then wrongly appear to him as its [(the mind's)] boundlessness or inexhaustibility.[23]

Moreover, "[it] would in no way derogate the incompletability of objective mathematics"; "[o]n the contrary, it would only make it particularly striking."

> For if the human mind were equivalent to a finite machine, then objective mathematics not only would be incompletable in the sense of not being contained in any well-defined axiomatic system, but moreover there would exist *absolutely* unsolvable diophantine problems of the type described above, where the epithet "absolutely" means that they would

[21] He elaborates in a footnote: "For this (or the consequence concerning the consistency of the axioms) would constitute a mathematical insight not derivable from the axioms [and] rules under consideration, contrary to the assumption." *Ibid.*

[22] See also the conversational reports in Wang, *A Logical Journey*, *op. cit.*, section 6.1.7, p. 186 (quoted in section IV.3 below) and section 6.1.8, pp. 186–87.

[23] Gödel, "Some Basic Theorems on the Foundations of Mathematics and Their Implications," *op. cit.*, pp. 309–10.

> be undecidable, not just within some particular axiomatic system, but by *any* mathematical proof that the human mind can conceive.[24]

In other words, *if* there were an F^* such that $F^* = K$, *then* we would not only have $F^* \subsetneq T$ (and, more generally, $F \subsetneq T$ for any F such that $F \subseteq T$), but we would also have $K \subsetneq T$ (and hence there would be absolutely undecidable sentences, that is, sentences φ such that $\varphi \in T$ and yet neither $\varphi \in K$ nor $\neg\varphi \in K$).[25]

Gödel then reformulates this in disjunctive form.

> So the following disjunctive conclusion is inevitable: *Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, 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 of the type specified* (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives).[26]

In other words, we have *Gödel's Disjunction*:

> *Claim 3.* Either $\big(\neg\exists F\,(F = K)\big)$ or $\big(\exists\varphi\,(\varphi \in T \wedge \varphi \notin K \wedge \neg\varphi \notin K)\big)$.

The argument for the disjunction is not too hard to provide informally: Suppose that there were an F^* such that $F^* = K$. Then we have $F^* \subseteq T$ (since $K \subseteq T$). So, by the incompleteness theorems, we have $F^* \subsetneq T$. Thus, there is a $\varphi \in T$ such that $\varphi \notin F^*$, and for any such φ we also have $\neg\varphi \notin F^*$ (since if $\neg\varphi \in F^*$ then $\neg\varphi \in T$ (as $F^* \subseteq T$), which is impossible since $\varphi \in T$). But $F^* = K$. So this φ is such that $\varphi \in T$ and yet neither $\varphi \in K$ nor $\neg\varphi \in K$.

II. SHARPENING THE NOTIONS

Let us now start spelling out our assumptions on F, K, and T, and place the above informal discussion in a precise, formal setting, where we can establish definitive results. In this section we will introduce two systems of epistemic arithmetic—the first, EA, designed to deal with F and K, and the second, EA_T, designed to deal with F, K, and T.

[24] *Ibid.*, p. 310.

[25] If $K \subsetneq T$ then there must be a φ such that $\varphi \in T$ and $\varphi \notin K$, and for any such φ we also have $\neg\varphi \notin K$ (since if $\neg\varphi \in K$ then $\neg\varphi \in T$, which is impossible since $\varphi \in T$).

[26] Gödel, "Some Basic Theorems on the Foundations of Mathematics and Their Implications," *op. cit.*, p. 310, his italics.

II.1. Sharpening F. In the informal setting 'F' was used to stand for the set of sentences provable relative to a given formal system (or, in the variant formulation, the set of sentences that can be produced by "an idealized finite machine"). It is well-known how to sharpen this notion. In fact, in this case we have a *substantive* analysis of the notion: The informal notion of being "provable relative to a given formal system" is rendered mathematically precise in terms of the notion of being "provable relative to a recursive set of axioms."[27]

Now, all of this can be formalized in PA, which we will take as our base system. So, in all of the systems we will consider, we shall have at our disposal the resources to quantify over formal systems F_e (or, in the variant formulation, ideal finite machines, that is, Turing machines) in a perfectly precise manner. We will let 'F_e' stand for the set of sentences provable relative to the e^{th} recursively enumerable set of axioms, where 'e' ranges over the natural numbers.

II.2. Sharpening T. In the informal setting 'T' was used to stand for the set of sentences that are true. Here things are a little more delicate, since in this case it would be hard to do something comparable—it would be hard to give a substantive analysis of the notion of truth in more fundamental terms. However, we can hope to give a *structural* analysis; that is, instead of analyzing the notion in more fundamental terms we can hope to articulate the principles that capture its essential features.

For the purposes of the arguments that we will consider in this part of the paper—the first generation of arguments for the first disjunct—we will only require a *typed* truth predicate, and for typed truth we have a perfectly adequate structural analysis, namely that of Tarski.[28]

It would take us too far afield to list all of the Tarskian principles governing 'T'. Suffice it to say that these principles include such principles as

(T_1) $(\forall x)[\text{Sent}(x) \rightarrow (T(\dot{\neg}x) \leftrightarrow \neg T(x))]$ and
(T_2) $(\forall x)(\forall y)[\text{Sent}(x) \wedge \text{Sent}(y) \rightarrow (T(x \dot{\vee} y) \leftrightarrow T(x) \vee T(y))]$.

[27] This latter notion is provably co-extensive with the notion of "what can be produced by a Turing machine," and so we have simultaneously rendered precise the corresponding notion in the variant formulation.

[28] A *typed* truth predicate is one that applies only to statements that do not themselves involve the truth predicate. In contrast, a *type-free* truth predicate is one which also applies to statements that themselves involve the truth predicate. The principles governing typed truth predicates are perfectly straightforward and uncontroversial, while the principles governing type-free truth predicates are much more delicate.

In the successor to this paper—where we consider the second generation of arguments for the first disjunct—we will have to employ a type-free truth predicate.

The notation involving the dots is necessary for full precision, but the details will not be important for our purposes. The main point I wish to convey is that these principles capture uncontroversial aspects of the (classical) notion of truth. For example, T_2 simply says that a disjunction is true if and only if one of its disjuncts is true.

II.3. Sharpening K. In the informal setting '*K*' was used to stand for the set of sentences that are "absolutely provable" (or, in the variant formulation, the set of sentences that can be produced by "the idealized human mind"). There is little hope of giving a substantive analysis of this notion, as we did in the case of *F*. But perhaps we can retreat and, as in the case of *T*, provide a structural analysis.

The trouble is that, in contrast to the case of truth, there is little agreement even on what principles are supposed to govern the notion "absolute provability" (or the notion of what can be produced by "the idealized human mind") since there is little agreement on how absolute absolute provability is supposed to be (or how ideal the idealized human mind is supposed to be).

What we shall do is follow the charitable course. Our opponent—the proponent of the first disjunct—wishes to show that *K* outstrips any correct *F*. Without fully understanding the specific nature of *K*, and without even taking a stance on which principles are supposed to truly govern it, we will grant our opponent a very strong notion of *K* along with a powerful set of principles governing it. Notice that in doing so we are making the task for our opponent *easier*. For the more we grant our opponent concerning *K*, the easier the task of showing that *K* outstrips any *F*, and, correspondingly, the stronger any negative result we might establish to the effect that even such a strong notion of *K* cannot be shown to outstrip any *F*.

In the formal setting we will treat '*K*' as an operator.[29] The *basic axioms of absolute provability* are:

[29] In our present setting we are forced to treat '*K*' as an operator. This is because results of Gödel, Myhill, Montague, Thomason, and others show that under fairly general conditions (of which our present conditions are an instance) if one formulates a theory of absolute provability with '*K*' as a predicate then inconsistency ensues. See Kurt Gödel, "An Interpretation of the Intuitionistic Propositional Calculus" (1933), reprinted in *Collected Works, Volume I: Publications 1929–1936*, ed. Solomon Feferman et al. (New York: Oxford University Press, 1986), pp. 301–03; John Myhill, "Some Remarks on the Notion of Proof," this JOURNAL, LVII, 14 (July 1960): 461–71; Richard Montague, "Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability," *Acta Philosophica Fennica*, XVI (1963): 153–67; and Richmond H. Thomason, "A Note on Syntactical Treatments of Modality," *Synthese*, XLIV, 3 (July 1980): 391–95. (However, as we shall see in the successor to this paper, if one situates a theory of absolute provability within a *type-free* theory of truth, then one can treat '*K*' as a predicate and circumvent the aforementioned limitative results.)

(K_1) Universal closures of formulas of the form

$$K\varphi$$

where φ is a first-order validity.

(K_2) Universal closures of formulas of the form

$$(K(\varphi \to \psi) \wedge K\varphi) \to K\psi.$$

(K_3) Universal closures of formulas of the form

$$K\varphi \to \varphi.$$

(K_4) Universal closures of formulas of the form

$$K\varphi \to KK\varphi.$$

The first principle—known as *logical omniscience*—asserts that K holds of *all* first-order logical validities. The second principle asserts that K is closed under modus ponens, and so distributes across logical derivations. The third principle is a way of asserting that K is correct. And the fourth principle asserts that K is "absolutely self-reflective."

These principles indicate that we are indeed granting our opponent an extremely strong notion of K. To see this, consider the first principle. In the variant formulation this principle says that "the idealized human mind" knows all first-order logical validities. But notice that some (indeed, most) of the logical validities are too long for an actual agent to even comprehend, let alone know; in fact, some (indeed most) logical validities have more symbols *than there are fundamental particles in the observable universe*. So, in granting logical omniscience we are granting a strong notion of K, one that involves treating the "idealized human mind" in a highly idealized manner.[30]

These strong assumptions on K might seem like grand and questionable assumptions to make at the start of an attempt to make a case for the first disjunct. But I want to stress once again that the strong assumptions that I am making on K are made on behalf of my opponent. Since my goal is to show that my opponent's arguments do not establish the conclusion, the more I grant in terms of strong assumptions on K, the stronger any negative result to the effect that the arguments do not show that even such a strong notion of K does not coincide with any F.

[30] In the final section of the successor to this paper I will examine the nature of this idealization and conclude that already at this step there is an unjustified move. But for the time being I want to grant my opponent as much as possible and place more weight on limitative mathematical results than philosophical critique.

II.4. The Systems EA *and* $\mathrm{EA_T}$. We are now in a position to describe the systems that we shall be employing. The basic system EA of epistemic arithmetic has axioms of arithmetic and axioms of absolute provability, and the extended system $\mathrm{EA_T}$ has, in addition, axioms of (typed) truth.[31]

The language L_{EA} is L_{PA} expanded to include an operator 'K' that takes formulae of L_{EA} as arguments. The *axioms of arithmetic* are simply those of PA, only now the induction scheme is taken to cover all formulas in L_{EA}. For a collection Γ of formulas in L_{EA}, let '$K\Gamma$' denote the collection of formulas '$K\varphi$' where 'φ' is in Γ. The system EA is the theory axiomatized by $\Sigma \cup K\Sigma$, where Σ consists of the axioms of PA (in the language L_{EA}) and the basic axioms of absolute provability.

The language $L_{\mathrm{EA_T}}$ of $\mathrm{EA_T}$ is the language L_{EA} augmented with a unary predicate 'T'. The system $\mathrm{EA_T}$ is the theory axiomatized by $\Sigma \cup K\Sigma$, where Σ consists of the axioms of PA (in the language $L_{\mathrm{EA_T}}$), the basic axioms of absolute provability (in the language $L_{\mathrm{EA_T}}$), and the Tarskian axioms of truth (for the language L_{EA}).

III. GÖDEL REVISITED

Let us now return to Gödel's informal discussion, recasting it in the framework of $\mathrm{EA_T}$. It turns out that each of the three positive claims that Gödel makes (and which we described informally in section I) can be formalized and proved within $\mathrm{EA_T}$.

III.1. The First Claim. The first claim concerns the relationship between F and T. It asserts that for any formal system F,

$$F \subseteq T \rightarrow F \subsetneq T.$$[32]

This is formalizable and provable in $\mathrm{EA_T}$. But this is hardly surprising since, as we noted, this application of the incompleteness theorems is a straightforward meta-mathematical result.

[31] Systems of this kind were first introduced by Myhill, "Some Remarks on the Notion of Proof," *op. cit.*; William N. Reinhardt, "The Consistency of a Variant of Church's Thesis with an Axiomatic Theory of an Epistemic Notion," *Revista Colombiana de Matemáticas*, XIX, 1–2 (1985): 177–200; William N. Reinhardt, "Absolute Versions of Incompleteness Theorems," *Noûs*, XIX, 3 (September 1985): 317–46, William N. Reinhardt, "Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems," *Journal of Philosophical Logic*, XV, 4 (November 1986): 427–74; and Stewart Shapiro, "Epistemic and Intuitionistic Arithmetic," *Studies in Logic and the Foundations of Mathematics*, CXIII (1985): 11–46, and have been investigated by many others (see, for example, Leon Horsten, "In Defense of Epistemic Arithmetic," *Synthese*, CXVI, 1 (1998): 1–25; Hannes Leitgeb, "On Formal and Informal Provability," in Otávio Bueno and Øystein Linnebo, eds., *New Waves in Philosophy of Mathematics* (New York: Palgrave Macmillan, 2009), pp. 263–99, and the references therein).

[32] Now, in the setting of $\mathrm{EA_T}$, the notation '$F \subseteq T$' is being used as a convenient shorthand for '$\forall x \left(\mathrm{Sent}_{L_{\mathrm{PA}}}(x) \rightarrow (F_e(x) \rightarrow T(x))\right)$', where F_e is the e^{th} recursive set of axioms.

III.2. The Second Claim. The second claim involves K. It asserts that for any formal system F,

$K(F \subseteq T) \rightarrow F \subsetneq K$.

This is also provable in EA_{T}. In fact, something stronger is provable already in EA. The antecedent of the above conditional is the statement that it is absolutely provable that F is correct (that is, $K(F \subseteq T)$). But a weaker condition is the scheme asserting that for any φ in L_{PA}, it is absolutely provable that if F holds of φ then φ holds (that is, for any φ in L_{PA}, $K(F(\ulcorner\varphi\urcorner \rightarrow \varphi))$).[33] In order to express this weaker condition we do not require the truth predicate, since in this case we are dealing with the statements φ one by one (and not quantifying over them), and so we can replace '$T(\ulcorner\varphi\urcorner)$' with '$\varphi$'.

Theorem 3 (Reinhardt). Assume that S includes EA. Suppose $F(x)$ is a formula with one free variable and is such that for each sentence φ

$S \vdash K(F(\ulcorner\varphi\urcorner) \rightarrow \varphi)$.

Then there is a sentence ψ such that

$S \vdash K\psi \wedge K\neg F(\ulcorner\psi\urcorner)$.[34]

Notice that here '$F(x)$' stands for *any* formula in L_{EA} with one free variable. It need not be a formula defining what is provable relative to a recursive set of axioms; that is, it need not be a Σ^0_1-statement. The result applies to even richer notions of relative provability.

It is worth mentioning that one also obtains an absolute version of the second incompleteness theorem:

Theorem 4 (Reinhardt). Assume that S includes EA. Suppose $F(x)$ is a formula with one free variable and is such that for each sentence φ

$S \vdash K(K\varphi \rightarrow F(\ulcorner\varphi\urcorner))$.

Then

$S \vdash K\neg K(\mathrm{Con}(F))$.[35]

In other words, if it is absolutely provable (pointwise) that the system F captures everything that is absolutely provable then (it is absolutely provable that) the consistency of F is not absolutely provable.

[33] Here $\ulcorner\varphi\urcorner$ is the Gödel code for the statement φ.
[34] Reinhardt, "Absolute Versions of Incompleteness Theorems," *op. cit.*
[35] *Ibid.*

III.3. The Third Claim. The third claim is the disjunction, that is,

$$\big(\neg\exists F\,(F = K)\big) \vee \big(\exists\varphi\,(T(\varphi) \wedge \neg K(\varphi) \wedge \neg K(\neg\varphi))\big).$$

This is a little delicate to formalize in $\mathrm{EA_T}$ since K is formalized as an operator in $\mathrm{EA_T}$ and so we are prohibited from quantifying into it. The solution is to employ the truth predicate and use it to replace '$\exists x\, K(x)$' with '$\exists x\, T(\dot{K}x)$'. The metamathematical details (like the role of the dot under the 'K') are a bit tedious, but routine. Letting '$x \in F_e$' be shorthand for the statement "x is in the e^{th} recursively enumerable set," we can now express the first disjunct as

$$\neg\exists e\,\big(\forall x\,(\mathrm{Sent}_{L_{\mathrm{PA}}}(x) \rightarrow (T(\dot{K}x) \leftrightarrow x \in F_e))\big)$$

and we can express the second disjunct as

$$\exists x\,\big(\mathrm{Sent}_{L_{\mathrm{PA}}}(x) \wedge T(x) \wedge \neg T(\dot{K}x) \wedge \neg T(\dot{K}\dot{\neg}x)\big).$$

The formal statement of Gödel's Disjunction is the disjunction of these two statements, which we shall abbreviate as 'GD'.

Theorem 5 (Reinhardt). Assume $\mathrm{EA_T}$. Then GD holds.[36]

III.4. Summary. Thus, once the background assumptions on the fundamental concepts—F, K, and T—are made explicit, the entire discussion can be pulled into a framework—$\mathrm{EA_T}$—in which one can prove definitive results. Remarkably, as we have seen, each of Gödel's three main claims is provable in $\mathrm{EA_T}$. In this sense, Gödel was correct in claiming that the disjunction is a "mathematically established fact."

So the question arises: Which disjunct holds? In particular, can we go further and establish that the first disjunct is a "mathematically established fact"?

IV. LUCAS AND PENROSE: THE FIRST DISJUNCT

The first generation of arguments for the first disjunct—due primarily to Lucas[37] and early Penrose[38]—has been discussed extensively in the literature. In our present setting, with the above technical apparatus at hand, we can both sharpen the debate and present a critique that is a good deal stronger than the standard critiques.

[36] Reinhardt, "Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems," *op. cit.*

[37] Lucas, "Minds, Machines and Gödel," *op. cit.*

[38] Penrose, *The Emperor's New Mind*, *op. cit.*

IV.1. The Classic Argument for the First Disjunct. It turns out that the first generation of arguments for the first disjunct are really just versions of Gödel's argument for his second claim, namely that *if* $K(F \subseteq T)$ then $F \subsetneq K$. In the words of Penrose: "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth."[39] As we saw above, this conclusion, when suitably formalized, is provable in $\mathrm{EA_T}$. (See Theorem 3 and the discussion surrounding it.)

However, as Gödel (and many logicians who followed) pointed out, the argument does not yield the first disjunct; rather, it provides us with a *conditional* statement, and to arrive at the consequent of the conditional, one needs to discharge the antecedent; that is, one needs the additional premise that $K(F \subseteq T)$.

The question, then, is whether for *any* F one can determine (in the sense of absolutely prove or refute) whether or not F is correct. This would involve, in the very least, being able to determine whether or not F is consistent. But this is no small task. For example, let S be the system PA + R where 'R' stands for the famous open problem known as the Riemann Hypothesis. A result of Kreisel shows that R can be formulated as a Π^0_1-sentence. It follows that PA + R is consistent if and only if R is true.[40] So to know whether or not PA + R is consistent is to know whether or not R is true. But R is a major outstanding problem in mathematics, so outstanding that the Clay Institute has offered one million dollars for its resolution. No one at present knows the answer to R, and it is no small task to determine the answer. It follows that no one at present knows whether or not PA + R is consistent.

Now, one might push back on this argument by pointing out that although Lucas and Penrose do not know whether PA + R is consistent, they might plausibly maintain that the answer is indeed within the reach of what is absolutely provable. For after all, Lucas and Penrose are, like the rest of us, actual humans, with the limitations and defects that come from being finite, real-world beings; but we are not here concerned with the performance of actual human minds, we are concerned with what the *idealized* human mind can do *in principle*. So Lucas and Penrose might plausibly maintain that although the answer to the question of whether PA + R is consistent is not actually within

[39] Penrose, *Shadows of the Mind*, *op. cit.*, p. 76.

[40] The reasoning is as follows: We know that PA is Σ^0_1-complete, and we are assuming that PA is Σ^0_1-sound (indeed, in $\mathrm{EA_T}$ we have that K holds of PA, and since K outputs only truths this means that PA is correct). So, if PA + R is consistent, then R must be true (since if F were false, then, by Σ^0_1-completeness, we would have PA $\vdash \neg R$); and, if PA + R is inconsistent, then PA $\vdash \neg R$, and so, by Σ^0_1-soundness, R must be false.

reach, it is in principle within reach. But the choice of R was merely representative, and the point is much stronger: To know of *every* system F whether or not F is consistent (something that they maintain these idealized human minds are capable of doing) is to have an oracle for Π^0_1-truth, and that is not something one can claim to have at the *start* of an argument for the first disjunct, since it trivially contains the conclusion of the argument.

These considerations show that this *particular* argument for the first disjunct fails, but perhaps there is another argument . . .

IV.2. The First Disjunct in $\mathrm{EA_T}$. In fact, we can say something much stronger. Not only does this particular argument fail to establish the first disjunct, even granting the very strong notion of K that is embodied in $\mathrm{EA_T}$; there is *no* argument for the first disjunct in $\mathrm{EA_T}$.

To describe the limitative results and the subtle issues involved, it is useful to distinguish (following Reinhardt) three grades of the mechanistic thesis:

(1) (WMT) $\exists e\,(K = F_e)$
(2) (SMT) $K\,\exists e\,(K = F_e)$
(3) (SSMT) $\exists e\,K(K = F_e)$

The first thesis is the *weak mechanistic thesis.*[41] It asserts that there is a Turing machine which coincides with the idealized human mind (in the sense that the two have the same outputs). This is simply the first disjunct of Gödel's Disjunction. The second thesis is the *strong mechanistic thesis.* It asserts that the idealized human mind knows that there is a Turing machine which coincides with the idealized human mind. The third thesis is the *super strong mechanistic thesis.* It asserts that there is a particular Turing machine such that the idealized human mind knows that *that* particular machine coincides with the idealized human mind.

The first result that is of relevance to this discussion is the following:

Theorem 6 (Reinhardt). '$\mathrm{EA_T}$ + SSMT' is inconsistent.[42]

In other words, in the context of $\mathrm{EA_T}$, it is true that there cannot be a Turing machine such that the idealized human mind knows that it coincides with *that* machine.

This fact does not vindicate the proponents of the first disjunct. The statement that there is a Turing machine such that the idealized human mind knows that it coincides with *that* machine, is a rather strong

[41] In what follows I shall express matters in terms of the variant formulation.
[42] Reinhardt, "Absolute Versions of Incompleteness Theorems," *op. cit.*

statement. As we have seen, it is refutable in EA_T. But the mechanist is not maintaining such a strong statement. The mechanist is maintaining the more modest statement that there is a Turing machine that coincides with the idealized human mind; that is, the mechanist is maintaining WMT. *This* is what the proponents of the first disjunct are denying, and it is what they are claiming to have refuted on the basis of the incompleteness theorems. But now the following result comes into play:

Theorem 7 (Reinhardt). 'EA_T + WMT' is consistent.[43]

In other words, from the point of view of EA_T it is entirely possible that the idealized human mind is in fact a Turing machine. It just cannot know which one.[44] This shows that there is *no* argument for the first disjunct in EA_T, and since EA_T would seem to embody all of the assumptions held by the proponents of the first disjunct it shows that there is a fundamental obstacle.

In fact, there is an even stronger conclusion. Reinhardt conjectured that even SMT is consistent with EA_T, and Carlson proved this conjecture, via a sophisticated construction:

Theorem 8 (Carlson). 'EA_T + SMT' is consistent.[45]

In other words, from the point of view of EA_T it is entirely possible that the idealized human mind *knows* that it is a Turing machine. It just cannot know which one.

These results show that if one is to have a hope of establishing the first disjunct, one must either invoke stronger assumptions or shift to an entirely new framework.

IV.3. A Subtle Distinction. There is a subtle distinction that is likely to have led people astray in thinking that they possessed a proof of the first disjunct.

The distinction was glimpsed by Gödel in his discussion of his third main claim. Here is another quote from Wang's reports on his conversations with Gödel:

> The incompleteness results do not rule out the possibility that there is a theorem-proving computer [F_e] which is in fact equivalent to mathematical intuition [$K = F_e$]. But they imply that, in such a—highly unlikely

[43] Reinhardt, "The Consistency of a Variant of Church's Thesis with an Axiomatic Theory of an Epistemic Notion," *op. cit.*

[44] This result gives precise mathematical substance to the possibility raised by Gödel (see section I.5 above and section IV.3 below) and later raised by Benacerraf in "God, the Devil, and Gödel," *op. cit.*

[45] Timothy J. Carlson, "Knowledge, Machines, and the Consistency of Reinhardt's Strong Mechanistic Thesis," *Annals of Pure and Applied Logic*, CV, 1–3 (2000): 51–82.

> for other reasons—case, either we do not know the exact specification of the computer [$\neg K\,(K = F_e)$] or we do not know that it works correctly [$\neg K(F_e \subseteq T)$].[46]

This can all be made precise and rigorous in the setting of $\mathrm{EA_T}$: The first sentence is substantiated by Reinhardt's result (Theorem 6) that $\exists e\,(K = F_e)$ is consistent with $\mathrm{EA_T}$. The second sentence has two parts. The first part is substantiated by Reinhardt's result (Theorem 7) that $\mathrm{EA_T}$ can prove $\neg\exists e\,K(K = F_e)$, and the second part is substantiated by the result (see Theorem 3) that Gödel's second claim is provable in $\mathrm{EA_T}$.

It is striking that in addition to being correct about his three main positive claims Gödel was able to appreciate these subtleties and glimpse both Theorem 6 and Theorem 7. He in effect anticipated every move in the subsequent debate and he was able to avoid the pitfalls that beset others.

The proponents of the first disjunct may have seen an informal argument for the fact that the incompleteness theorems imply that $\neg\exists e\,K(K = F_e)$. But it does *not* follow that $\neg\exists e\,(K = F_e)$, as Theorem 7 demonstrates. In fact, it does not even follow that $\neg K\exists e\,(K = F_e)$, as Theorem 8 demonstrates. The difference between '$\exists e\,K$' and '$K\exists e$' before '$(K = F_e)$' is paramount. It is possible (as far as the principles embodied in $\mathrm{EA_T}$ are concerned) to "know that you are a Turing machine" ($K\exists e\,(K = F_e)$); it is just not possible for there to be a Turing machine and "know that you are *that* Turing machine" ($\exists e\,K\,(K = F_e)$).

Regardless of what may have led Lucas and Penrose astray, the above discussion shows that the arguments for ¬WMT are not valid and, moreover, that there is *no* valid argument for ¬WMT (or even ¬SMT) that proceeds on the basis of $\mathrm{EA_T}$ alone. If one is to hope to prove the first disjunct, one must either invoke additional assumptions or shift to an entirely new framework.

V. CONCLUSION

One of the main things I have tried to illustrate is that these questions can be approached with precision by making the underlying assumptions governing the fundamental concepts explicit. We saw that Gödel's informal arguments for his three central philosophical claims—most notably, the disjunction—could be rendered formally precise and proved in the system $\mathrm{EA_T}$, thereby vindicating, in some

[46] Wang, *A Logical Journey, op. cit.*, section 6.1.7, p. 186.

sense at least, his claim that the disjunction is a "mathematically established fact." In this setting we could also clearly isolate the problems with the early arguments for the first disjunct and, more importantly, show that there is no argument for the first disjunct within EA_T, a system that would seem to cover every assumption that the proponents of the first disjunct would be willing to make.[47] I hope that this puts to rest the first generation of arguments for the first disjunct and that all participants in the dispute can agree on this.

The question of whether the second generation of arguments—that is, Penrose's new argument—establish the first disjunct is quite subtle. Recall that Gödel had hoped that when we had an adequate resolution of the paradoxes—most notably an adequate type-free theory of truth and absolute provability—we would be in a position to establish the first disjunct. We now have many type-free theories of truth. And it turns out that to formalize Penrose's new argument one must employ a type-free theory of truth. So perhaps Penrose has fulfilled Gödel's hope. The entire argument can be made precise, and, when one does this, something interesting emerges. But that's another story. . .

PETER KOELLNER

Harvard University

[47] I have said that EA_T would *seem* to cover every such assumption and not that it *does* cover every such assumption, since the proponents of the first disjunct do not clearly state all of their assumptions on the fundamental concepts. In any case, EA_T covers the assumptions that they *do* make explicit, and it is hard to see what kind of implicit assumptions going beyond EA_T are at play in their arguments.