WITTGENSTEIN, TURING, AND THE “FINITUDE” OF LANGUAGE

PAUL M. LIVINGSTON
pmliving@unm.edu
University of New Mexico

ABSTRACT. I consider the sense in which language is “finite” for Wittgenstein, and also some of the implications of this question for Alan Turing’s definition of the basic architecture of a universal computing machine, as well as some of the vast technological, social, and political consequences that have followed from it. I shall argue that similar considerations about the relationship between finitude and infinity in symbolism play a decisive role in two of these thinkers’ most important results, the “rule-following considerations” for Wittgenstein and the proof of the insolubility of Hilbert’s decision problem for Turing. Fortuitously, there is a recorded historical encounter between Wittgenstein and Turing, for Turing participated in Wittgenstein’s “lectures” on the foundations of mathematics in Cambridge in 1939; their interactions are documented in the text *Wittgenstein’s Lectures on the Foundations of Mathematics* edited by Cora Diamond.[1] Although my aim here is not to adduce biographical details, I think their exchange nevertheless evinces a deep and interesting problem of concern to both. We may put this problem as that of the relationship of language’s finite symbolic *corpus* to (what may seem to be) the infinity of its meaning.

I

Wittgenstein’s philosophy of mathematics has sometimes been described as finitist; but, as I shall argue here, his actual and consistent position on the question of the finite and infinite in mathematics and language is already well expressed by a remark in his wartime *Notebooks*, written down on the eleventh of October, 1914: “Remember that the ‘propositions about infinite numbers’ are all represented by means of finite signs!” (Wittgenstein 1979, p. 10) The point is neither that signs cannot refer to infinite numbers nor that propositions referring to them are meaningless or somehow otherwise out of logical order. It is, rather, that *even* propositions referring to infinite numbers – for instance the hierarchy of transfinite cardinals discovered by Cantor – must *have* their sense (and hence their capability to represent 'infinite quantities') by and through a finite symbolization, for instance through a proof of finite length. That is, it must be such a proof – given in a finite number of steps and stated in a language with a finite number of symbol types – that gives us whatever epistemic access we can have to infinite quantities and numbers. This is closely connected with the remark, made several times in the *Notebooks*, that is also destined to serve as a kind of *leitmotif* underlying the *Tractatus*' discussion of analysis, showing, and the elusive nature of logical form: that logic must "care for itself." Here, this means that all the forms of possible meaning must already show up in the (formal) possibilities of signification in a finite, combinatorial language. Wittgenstein concludes the entry for the eleventh of October by noting: "The propositions dealing with infinite numbers, like all propositions of logic, can be got by calculating the signs themselves (for at no step does a foreign element get added to the original primitive signs). So here, too, the signs must themselves possess all the logical properties of what they represent" (pp. 10–11). Thus, the problem of the meaning of the infinite is a problem of the *logic* or *grammar* of finite signs – of how, in other words, the (formal) possibilities of signification in a finite, combinatorial language can give us whatever access we can have to infinite structures and procedures.

In the lectures delivered in Cambridge in 1939, Wittgenstein proposes to discuss the "foundations of mathematics," but not in order either to contribute to the analysis and description of such foundations or to give new calculations or even interpretations of calculations in mathematics itself.[2] Rather, his aim is to remove certain misinterpretations or confusions that surround the analysis of the "foundations of mathematics," particularly with respect to what is involved in the understanding and meaning of mathematical structures.[3] Wittgenstein emphasizes that in speaking of understanding a mathematical structure, for instance a regular series of numbers or indeed the sequence of counting numbers themselves, we may speak of coming to "understand" the sequence; we may also speak of gaining a capability or mastering a "technique." Yet what it is to "understand" (to "know how to," or "to be able to," continue "in the same way") is not clear. The issue is the occasion for Turing's first entrance into the discussion, in lecture number II:

> *Wittgenstein*: We have all been taught a technique of counting in Arabic numerals. We have all of us learned to count – we have learned to construct one numeral after another. Now how many numerals have you learned to write down?
>
> *Turing:* Well, if I were not here, I should say $\aleph_0$.
>
> *Wittgenstein*: I entirely agree, but that answer shows something.
>
> There might be many answers to my question. For instance, someone might answer, "The number of numerals I have in fact written down." Or a finitist might say that one cannot learn to write down more numerals than one does in fact write down, and so might reply, 'the number of numerals which I will ever write down.' Or of course, one could reply " $\aleph_0$", as Turing did.
>
> Now should we say: "How wonderful – to learn $\aleph_0$ numerals, and in so short a time! How clever we are!"?—Well, let us ask, "How did we learn to write $\aleph_0$ numerals?" And in order to answer this, it is illuminating to ask, "What would it be like to learn only 100,000 numerals?"…
>
> I did not ask "How many numerals are there?" This is immensely important. I asked a question about a human being, namely, "How many numerals did you learn to write down?" Turing answered "$\aleph_0$" and I agreed. In agreeing, I meant that that is the way in which the number $\aleph_0$ is used.
>
> It does not mean that Turing has learned to write down an enormous number. $\aleph_0$ is not an enormous number.
>
> The number of numerals Turing has written down is probably enormous. But that is irrelevant; the question I asked is quite different. To say that one has written down an enormous number of numerals is perfectly sensible, but to say that one has written down $\aleph_0$ numerals is nonsense. (Diamond 1976, p. 31)

Notably, Wittgenstein does not, here, *at all* deny the validity of the response that Turing initially (if guardedly) offers to the question about the capacity to write down numbers. Indeed, in endorsing Turing's answer he distinguishes himself quite clearly from the finitist who would hold that the grammar of "can" goes no farther than that of "is," that I cannot justifiably say that my capacity includes any more than actually has occurred or will occur. In knowing how to write down Arabic numerals, a capacity we gain at an early age and maintain throughout our rational lives, we possess a capacity that is rightly described as the capacity to write down $\aleph_0$ different numbers. The attribution of this capacity is not, moreover, an answer to the "metaphysical" question of how many numbers there *are*; the question is, rather, what *we*, as human beings possessing this familiar capacity, are thereby capable *of*.

Yet how is this recognizably infinitary capacity underlain by our actual contact, in learning or communication, with a finite number of discrete signs (or sign-types) and a finite number of symbolic expressions of the rules for using them? It is not difficult to see this as the central question of the so-called "Rule-Following Considerations" of the *Philosophical Investigations*, some of which was already extant in manuscript by 1939 (see, e.g., *PI* 143–155; 185–240). However, we may also, I think, see this very question as *already* decisive in Turing's development of the definition of a "universal computing machine" and its application to demonstrate the unsolvability of Hilbert's decision problem in the remarkable "On Computable Numbers, with an Application to the Entscheidungsproblem" written three years earlier, in 1936, and published in 1937. Turing's aim is to settle the question of whether there are numbers or functions that are not computable; that is, whether there are real numbers whose decimals are not "calculable by finite means" (Turing 1936, p. 58). He reaches the affirmative answer by defining a "computing machine" that works to transform given symbolic inputs, under the guidance of internal symbolic "standard descriptions," into symbolic outputs.

According to what has come to be called "Turing's thesis," (or sometimes the "Church-Turing" thesis) every number or function that is "effectively" computable at all (in an "intuitive" sense of effective computability) is computable by some Turing machine, and thus that the architecture of the Turing machine indeed captures, replaces, or formalizes the "intuitive" notion of computability. The thesis is, today, almost universally accepted; however, this should not blind us to the depth of the philosophical issues involved in this particular way of understanding the nature of a technique or procedure and the kind of relation between a finite calculus and its (potentially) infinite application that it suggests. According to Turing's thesis, for instance, what it is for anything (function or number) to be calculable at all is for it to be calculable by "finite means" (here, using only a finite number of lexicographically distinct symbols and finitely many symbolically expressible rules for their inscription and transformation). Twice in the article (p. 59 and pp. 75–76), Turing justifies these restrictions by reference to the finitary nature of human cognition, either in memory or in terms of the (necessarily finite) number of possible "states of mind";[4] similarly, he supposes that we can distinguish between at most finitely many "mental states;" accordingly, it is necessary that a Turing machine can have only finitely many distinct states or operative configurations, and that its total "program" can be specified by a finite string of symbols.

These restrictions prove fruitful in the central argument of "On Computable Numbers," to show that there are numbers and functions that are *not* computable in this sense. The first step is to show how to construct a *universal* Turing machine, that is, a machine which, when given the standard description of any particular Turing machine, will mimic its behavior by producing the same outputs (pp. 68–69). Because each standard description is captured by a *finite* string of symbols, it is possible to enumerate them and to work with the numbers (Turing calls them "description numbers") directly (pp. 67–68). Given that we know how to construct a universal machine, we now assume for *reductio* that there is a machine, H, that will test each such description number to determine whether it is the description number of a machine that halts when given its own description number as an input (p. 73).[5] It does this by simulating the behavior of each machine when it is given its own description number as an input. We also know that H itself, since it always produces a decision, always halts. However, the machine H itself has a description number, K. Now we consider what happens when the hypothesized machine considers "itself," that is evaluates whether the machine corresponding to the description number K halts. We know by hypothesis that the machine H halts; however, as Turing shows, it cannot. For in considering K, the machine enters into an unbreakable circle, calling for it to carry out its own procedure on itself endlessly. We have a contradiction, and therefore must conclude that there can be no such machine H.[6]

The result at the heart of Turing's paper is thus an application of the general formal or metalogical procedure, first discovered by Cantor, known as "diagonalization." The procedure underlies Cantor's own identification of the transfinite cardinals, as well as Gödel's two incompleteness theorems. Gödel's application of a procedure of “artithmetizing” syntax is, indeed, quite similar to Turing’s numbering of the Turing machines; and as Turing points out, Gödel’s first theorem is indeed itself an implication of his own result.[7] It is no accident that both of these decisive results of metalogic rely on what Turing calls an application of the “diagonal procedure,” in which the enumerability of syntax (here: of standard descriptions) is the key to the possibility of an application of the regular structure of a symbolic system to “itself,” and hence to produce a particular local configuration (the Gödel sentence or Turing’s machine H) that stands, almost paradoxically, both within and without the system whose logic it captures. As Graham Priest (2002) has recently argued, the general structure of diagonalization can in fact be seen as underlying an exceedingly wide variety of problematic and paradoxical results in the history of philosophy, whenever theoretical reflection grasps the limits of thought in their (paradoxically thinkable) determinacy. With respect to language, this is equivalent to the attempt, common to Gödel and Turing, to model the formal capabilities of a system within that system itself, by way of arithmetization or enumeration. It is in this way that the determining syntax of the system – the rules determinately responsible for all of its capabilities – are captured and metalogically reflected “back” into the system itself, producing the point of undecidability or indeterminacy.

The basis for this possibility in the results of Gödel and Turing alike is the possibility of “numbering” symbolic strings and so intervening on them. In this respect, one can say that diagonalization (whatever else it may be) is *always an intervention on symbolic expressions*; that is, it depends in a decisive way on the fact that meaningful procedures are *necessarily* captured, if at all, in a combinatorial symbolic expression that itself combines one or more signs according to definite rules. That is, diagonalization is in each case an intervention, not on procedures or numbers themselves, but on the *ways* procedures and numbers are necessarily expressed by means of finite strings of finitely many distinct symbols. This syntactical reference is essential for all forms of diagonalization, and it may thus be seen that the possibility of diagonalization and its results depends *essentially*, in each case, on the fact that language must make use of finitary means – a finite stock of symbols and a finite expression of rules – to accomplish its infinitary powers of symbolization.[8]

Now, it is familiar that Wittgenstein held, in general, a dim view of the purported *results* of various forms of the "diagonal procedure," including both Cantor's multiple infinites and the truth of Gödel's "self-referential" sentence. Do these doubts, expressed prominently in the *Remarks on the Foundations of Mathematics*, imply that there is not a very similar concern about the relationship of finite symbolism to infinitary techniques operative in Wittgenstein's own thought about rules and symbols? I think not, for the following reasons. In his critical remarks about the Gödel sentence as well as about Cantor's multiple infinities, Wittgenstein emphasizes that the existence of a procedure – even one with no fixed end, like the procedure of writing down numbers in Arabic numerals – does not imply the existence of a superlative *object*, either a "huge number" or a completed list of decimal expansions that itself contains "infinitely many" members. To a certain extent at least, these suspicions extend to the "diagonal procedure" itself. Though Cantor can, with some justice, say how one *can* generate a decimal expansion that, as one can show, does not appear anywhere on an "infinite list" of expansions, he has not *in fact* generated it; diagonalization is always in fact the "outcome" of an infinite procedure and cannot be said to have finished. However, Wittgenstein does not deny that there *is* such a procedure, and even that we can speak of it, with some justice, as one that shows (by giving sense to the proposition) that there is, for any set of decimal expansions, one that is not in this set (*RFM* II-29). Cantor has given us a procedure that allows us to say: *given* any series of numerical symbols, we *can* (i.e. we have a method that lets us) generate a different one. However, in understanding the possibility and implications of this procedure, we must also keep in mind that there is a difference between series of numerical *symbols* and series of *numbers* in the mathematical sense. *A series in the mathematical sense is not a sequence of signs but a method for generating sequences of signs*.[9] There are analogies between the two uses, but they are different; and given the difference, Wittgenstein suggests, the existence of a sign ("$\aleph_0$") that expresses the unlimited possibility – the unlimitedness of the method – of generating sequences of signs does not by itself ground a further calculus with this sign, for instance one relating it to "other" infinities or other sizes of infinity. Nevertheless, as we have seen, it is just this ambiguity between sequences of signs and methods for generating sequences of signs upon which the claim of diagonalization to establish "positive" results depends. Diagonalization intervenes upon what are in fact sequences of signs (series in the non-mathematical sense) to produce a new number, a new sequence of signs which may itself be unlimited. What operates in this ambiguity, and creates the "crossing" at infinity (real or illusory) between procedures and their symbolization that is essential to diagonalization, is our presumed *infinitary* capacity to produce symbols according to well-defined rules.

In adducing these distinctions and casting doubt on the positive results of diagonalization, Wittgenstein's point is emphatically not, however, to show the nonexistence or invalidity of diagonalization as an (infinitary) *technique*. Rather, it is to emphasize the extent to which this procedure or technique, as infinitary as it is, has a place within a human life, and does not derive its meaning or sense from any other source than this life itself. Much later, in *RFM*, Wittgenstein comes back to this point:

> The concept of the rule for the formation of an infinite decimal is – of course – not a specifically mathematical one. It is a concept connected with a rigidly determined *activity* in human life. The concept of this rule is not more mathematical than that of: following the rule. Or again: this latter is not less sharply defined than the concept of such a rule itself.—For the expression of the rule and its sense is only part of the language-game: following the rule. (*RFM* VII – 42, p. 409)

Again, Wittgenstein is not here denying that there is a valid concept of the rule for the formation of an infinite decimal; nor that this rule is a rule for the formation of something that is indeed infinite. He is, rather, affirming that this formation – even in its strictness and rigidity – necessarily takes place as part of a human life, and gains its meaning and sense from this life. As it is capable of such infinite results, it would not, it seems, be quite right to call such a life, or the practice of following a rule within it (the language-game) that brings these about, "finite." Rather, the practice is *precisely* a technique: something of which beings with a finite spatiotemporal extent are capable, but whose *extension* is in principle without limit. It is thus neither the finitude of language nor the infinitude of meaning that makes possible its effect, but rather the gulf between them, in which Wittgenstein recognizes the openness of a human life.

There are, I think, two preliminary conclusions that can be drawn so far. The first is exegetical: Wittgenstein was certainly not in 1939, and probably never was, a finitist. That is, he *never* held that the finite character of language implied the non-existence or non-reality of infinite procedures. Rather, his focus is uniformly on the problem of the *grammar* of the infinite procedure: that is, just *how it is* that finite signs handled by finite beings gain the sense of infinity. This is none other than the radically posed question of the later Wittgenstein's thought: the question of the nature of a technique or practice. And it leads to the second conclusion, which is not exegetical but philosophical: that the infinity of technique is not an extension or intensification of the finite; nor is it a superlative or transcendent object that lies "beyond" all finite procedures. The infinity of technique enters a human life, rather, at the point of what might seem at first a radical paradox: that of its capture in finite signs, the crossing of syntax and semantics wherever the infinite rule is thought and symbolized as finite.

II

Given this suggestion of a rather close connection between the implication of diagonalization and the upshot of Wittgenstein's own rule-following considerations, how should we indeed view the sharply critical attitude he takes, both throughout the *Remarks on the Foundations of Mathematics* and elsewhere, toward Gödel's incompleteness theorems themselves (surely to be reckoned among the most important outcomes of the "diagonal procedure")? These remarks (where they have not been assumed to show that Wittgenstein simply "misunderstood" Gödel's result) have often been taken as support for an interpretation of his philosophy of mathematics as finitist or intuitionist, in that they have been taken as resting on a finitist denial of the utility or possibility of the "diagonal procedure." But although it is true that, as Wittgenstein reminds us, diagonalization is *essentially* an infinite procedure, he does not, as we have seen, deny its existence or possible utility. Moreover, in considering his response to Gödel, we ought to keep in mind Wittgenstein's remark in *RFM* that his purpose is not to address Gödel's proof (that is, presumably, not to affirm *or* deny it) but rather to "bypass it" (*RFM* VII, sect. 19). In particular, as Floyd and Putnam (2000) have recently argued, close attention to Wittgenstein's most notorious remarks about Gödel's proof shows that his point is not at all to deny Gödel's formal proof, but rather to suggest alternative possibilities for its interpretation. Here is the most crucial portion of these remarks:

> I imagine someone asking my advice; he says: "I have constructed a proposition (I will use 'P' to designate it) in Russell's symbolism, and by means of certain definitions and transformations it can be so interpreted that it says: 'P is not provable in Russell's system'. Must I not say that this proposition on the one hand is true, and on the other hand is unprovable? For suppose it were false; then it is true that it is provable. And that surely cannot be! And if it is proved, then it is proved that it is not provable. Thus it can only be true, but unprovable."
> Just as we ask; "'provable' in what system?", so we must also ask, "'true' in what system?" 'True in Russell's system' means, as was said: proved in Russell's system; and 'false in Russell's system' means: the opposite has been proved in Russell's system. — Now what does your "suppose it is false" mean? *In the Russell sense* it means 'suppose the opposite is proved in Russell's system'; *if that is your assumption*, you will now presumably give up the interpretation that it is unprovable. And by 'this interpretation' I understand the translation into the English sentence.—If you assume that the proposition is provable in Russell's system, this means it is true *in the Russell sense*, and the interpretation "P is not provable" again has to be given up. If you assume that the proposition is true in the Russell sense, *the same* thing follows. Further: if the proposition is supposed to be false in some other than the Russell sense, then it does not contradict this for it to be proved in Russell's system. (What is called "losing" in chess may constitute winning in another game.) (*RFM* I, Appendix III, sect. 8, pp. 118-19)

As Floyd and Putnam emphasize, although Wittgenstein does not dispute the validity of Gödel's proof itself, he raises the *question* of its correct interpretation. This does *not* involve disputing any of the mechanics that leads to the derivation of the "Gödel sentence" which "asserts" its own "unprovability." It *does* involve, however, raising a series of questions for the usual interpretation of the Gödel sentence that began with Gödel himself and has continued to be presupposed in most discussions of it. On this interpretation, the sentence shows the existence of a "mathematical truth" that cannot be proven by a formal system such as *Prinicipia Mathematica* and thus demonstrates the *incompleteness* of that system.[10]

Although this interpretation is still presupposed in virtually all discussions of Gödel's proof, it is reached, as Gödel himself pointed out, only through an essentially *informal* argument. (The argument is that *P* must be true, since if it were false "it would be true" that it can be proven, which cannot be the case, assuming the soundness of *PM*; and that since it can thus not be proven, and this is just what it "asserts," it is therefore true).[11] And although countless interpreters have followed Gödel in seeing his result as demonstrating the capacity of the human mind to grasp truths unprovable in any formal system, there *is,* as Floyd and Putnam point out, an alternative interpretation suggested by Wittgenstein's remarks. On this alternative, there is not (or at least there has not been shown to be) a unified sense of "truth" that subsumes the use of this predicate both *within* the formalism of *Principia Mathematica* and in the ordinary language in which the *informal,* metalogical argument is given. If we relax this assumption of a unified sense of "truth" between intra- and extra-systematic contexts, then we can see Gödel's formal result as having quite a different significance than Gödel himself suggests.

Specifically, recall that Gödel's first theorem constructs a sentence *P* such that, as is provable in *PM* or a related system, $P \leftrightarrow$ ~Prov([*P*]), where *Prov* is a one-place "provability predicate" and enclosure in square brackets gives the Gödel number of the formula enclosed. Additionally, the "provability predicate" itself is defined by means of the predicates NaturalNo(*x*), and Proof(*x,t*)*,* where NaturalNo(*x*) is interpreted as "*x* is a natural number" and Proof(*x,t*) is interpreted as a relation supposed to hold between two numbers when *x* is the Gödel number of a proof whose last line has the Gödel number *t.*[12] (Here, *t* abbreviates an expression which calculates out to the Gödel number of *P* itself). All of these are, of course, interpretations, and might be resisted under the right circumstances. *In particular,* suppose we actually assume that ~*P* is proven in *PM* (or, one day, actually come across a proof of it). Then we are in a position, of course, also to prove Prov([*P*]). In this case, however, as Wittgenstein points out, we might well be justified in dropping the *interpretation* that holds that Prov([*P*]) is *in fact* a provability predicate. And if we drop this interpretation, there is no need to conclude that the Gödel sentence is indeed something that is "true, but unprovable in *PM*."

How, though, might we justifiably drop the interpretation of Prov([*P*]) as "*P* is provable in *PM*"? As Floyd and Putnam point out, we might take the (successful, as we are now supposing) proof of ~*P* to demonstrate that *PM* is (not inconsistent but) ω-inconsistent.[13] If *PM* is ω-inconsistent, though, then in every admissible interpretation of *PM* (i.e., every interpretation which fits at least one model) there are, in addition to the natural numbers, entities which are *not* natural numbers; and NaturalNo(*x*) can no longer be interpreted as "x is a natural number." Moreover, Proof (*x*,*t*) can no longer be interpreted as relating the Gödel numbers of two formulas (one of which is a proof of the other), since in every admissible model its extension will contain some elements that are not natural numbers at all. This means that – supposing that there is a proof of ~*P* – it would no longer be tenable to *interpret* the Prov(x) predicate, defined in terms of the Proof (x,t) and NaturalNo(x), as "*P* is provable in *PM*." We would have to, as Wittgenstein suggests, "give up" this interpretation, and *along with it,* give up the interpretation of *P* as *saying* that it, itself, is unprovable.

Accordingly, Floyd and Putnam argue, it is in fact not possible simply to *assume* the informal interpretation that Gödel gave to his own theorem, that of showing the existence of "mathematical truths" that cannot be proven or disproven in any given system such as *PM*. As Wittgenstein effectively points out, we must distinguish here between what is actually established by the mathematical result itself, and the "metaphysical" claims that are made on its behalf:

> That the Gödel theorem *shows* that (1) there is a well-defined notion of "mathematical truth" applicable to every formula of *PM*; and (2) that, if *PM* is consistent, then some "mathematical truths" in *that* sense are undecidable in *PM,* is *not* a mathematical result but a metaphysical claim. But that if P is provable in *PM* then *PM* is inconsistent and if ~P is provable in *PM* then *PM* is ω-inconsistent is precisely the mathematical claim that Gödel proved. What Wittgenstein is criticizing is the philosophical naivete involved in confusing the two, or thinking that the former follows from the latter. But not because Wittgenstein wants simply to deny the metaphysical claim; rather, he wants us to see how little sense we have succeeded in giving it.[14]

More generally, at the heart of Wittgenstein's critical remarks about Gödel's proof is his skepticism that there is such a well-defined notion of "mathematical truth" that can be held in common between a system such as *Principia Mathematica* and the English "translations" of various of its notions, and so can license the usual interpretation of Gödel's result as showing that there are "truths" that cannot be proven in *Principia* (or any given system). In particular, if, as Wittgenstein suggests, there is indeed no *neutral* sense of "truth" that can be used to characterize both sentences in *PM* and their English translations, then there is no reason to suspect that Gödel's proof indeed shows what it has most often been taken to, that there is a "truth" that cannot be proven or disproven by *PM*. What we have, instead, is simply a particular sentence in *PM*, one that formulates a "perfectly ordinary" and undistinguished arithmetical claim, one that bears literally no implications for the powers or structure of the system as a whole.

When Gödel's theorem and its broader philosophical implications are discussed, the usual framework of discussion is a *model-theoretic* conception of truth. That is, the truth of the Gödel sentence *P* is conceived as a matter of its holding for a (natural) model, where it is assumed furthermore that there is at least one model where all of the objects of which it holds are natural numbers. As we have just seen, even remaining within a model-theoretic conception of truth, this last assumption is disputable, and might indeed well be disputed if a proof of ~*P* were to be given. However, just as importantly, the model-theoretic conception of truth itself might be disputed. Wittgenstein himself never held such a conception, tending to suggest instead a disquotational or redundancy theory.[15] On such a theory, as he suggests in the passage on Gödel's proof itself, there is no language- or system-independent notion of truth, and so there is no absolute sense to the claim that the Gödel sentence *P* expresses a "mathematical truth." Instead, as Wittgenstein suggests, the only available sense of "true" that is evidently applicable to the Gödel sentence, conceived as a sentence of *PM*, is the sense "proven in *PM*." Under the assumption that this is indeed the only relevant sense of "true," though, the Gödel sentence simply collapses to a version of the "Knower Paradox" (the sentence P that says: "P is known to be false") or the liar paradox: *P* iff it is not true that *P*.[16] (Here, we are still maintaining that *Prov*(x) can be interpreted as a

"Proof predicate" (and accordingly, under these assumptions, as a truth or knowledge predicate).)[17] This may again tend to suggest the inconsistency of *PM*, but crucially, it does not at all suggest that Gödel's proof bears witness to a substantial "truth" that is beyond the capacity of *PM* to prove.

To summarize, then, there are at least four ways, implicit in Wittgenstein's remarks, that we might resist the strong claim usually associated with Gödel's first incompleteness theorem (i.e. that it shows there is a "truth" that is beyond the capacity of *PM* to prove or disprove). First, we might simply abstain from interpreting the Gödel sentence *P* in terms of truth, falsity, provability, or "self-reference" at all. On this option, the derivability of the Gödel sentence in *PM* simply shows that a "perfectly ordinary" and unremarkable arithmetical sentence of *PM* is derivable. There are then, quite simply, no further consequences for the nature or structure of *PM* at all. Second, while agreeing to interpret the Gödel sentence in terms of issues of truth and provability, we might refuse the model-theoretic conception of truth and opt for a disquotational notion. Then the Gödel sentence is just equivalent to the Liar paradox, and raises the same issues as does that paradox. These may (but do not obviously) include the implication that *PM* is inconsistent.[18] Third, we might *agree* to both the interpretation in terms of truth and falsity and the model-theoretic conception of truth, and still resist the interpretation of "Prov(x)" as a "provability predicate"; this is the interpretation suggested by Floyd and Putnam, according to which there is no admissible interpretation of *PM* whose models do not contain objects that are not natural numbers, and *PM* is accordingly ω-inconsistent (although not necessarily inconsistent outright); and fourth (and finally), we may, on any of the first three options or for other reasons, take the Gödel sentence to show *PM* to be (outright) inconsistent.

On *any* of these four options, the Gödel sentence does not have the consequence of showing that "there is" a mathematical truth that can be neither proven nor disproven in *PM*. This is enough to underwrite Wittgenstein's marked suspicion about the result as it is usually presented, and to show that it would be over-hasty simply to concur with the metalogical interpretation that Gödel himself gives. It is not, in fact, completely clear which of these four "deflationary" options Wittgenstein himself favors; in his explicit remarks on Gödel's result he seems to waver among them. However, we may nevertheless draw some general conclusions from the availability of these four options itself. Significantly, on any but the first option, the Gödel sentence effectively *suggests* the inconsistency or at least ω-inconsistency of *Principia Mathematica* (or any system for which there is a Gödel sentence). This may seem, at first, an alarming suggestion, but note that this suggestion just amounts, in each case, to the suggestion that a sentence that has the "special" form of the Gödel sentence will produce a contradiction or antinomy; there is, as yet, no implication as to the further consequences or implications of such a contradiction.[19] And since the first option – on which the Gödel sentence is taken simply as a normal, arithmetical sentence of *PM* – seems to amount more to opting out of metalogic than pursuing it, we may well take this general implication of the other three options to be a generally legitimate one, assuming we wish both to interpret the Gödel sentence metalogically and resist the usual interpretation in terms of "incompleteness." Indeed, it seems we here have, once again, a vivid illustration of the general choice that the phenomena of systematicity and self-reference universally face us with: the choice between (consistency with) incompleteness (Gödel's interpretation, and the usual gloss on his result) *or* inconsistency and paradox (with the completeness of a system understood to be capable of formulating – though inconsistently! – its own logic of proof, or truth, entirely within itself).

It might seem at first as if this second way of looking at things is simply incoherent, or ruled out on metalogical grounds. Are we not in a position to *know* that *Principia Mathematica*, for instance, is not inconsistent, and so that it cannot contain the kind of contradiction that threatens to appear within it, on this interpretation? If the answer is indeed affirmative, then it might seem that we can rule out a Wittgenstein-style interpretation of Gödel's result and must indeed opt for the Gödel-style interpretation on which it demonstrates incompleteness. However, it is highly significant that we can "verify" the consistency of a system such as *Principia Mathematica* (for instance, by means of a model-theoretic soundness proof) *only from the position of a metalanguage* outside the system whose consistency is thereby proven. Moreover, Gödel's *second* "incompleteness" theorem shows precisely that the consistency of a system cannot be proven *by* that system itself. Thus, while we may be able to convince ourselves of the consistency of a system like *Principia Mathematica*, which we can step "outside of" and treat from the position of a metalanguage (here, English), where we are concerned with the *very* *system* we are ourselves using, we do not have this option.[20] In this case, it indeed becomes much more plausible that the constructability of the Gödel sentence for the system indeed implies the existence of contradictions, such as a sentence that says of "itself" that it is not (provable and hence not) true.

But does not the presence of such a contradiction vitiate the system in which we are working entirely, since (as can be formally shown) *anything* can be proven from a contradiction? The claim that it does, and hence the vehement desire to prohibit or rule out contradiction at virtually *any* cost, is one of the most prominent supports of the "foundationalist" picture of formal systems that is, in all of his engagements with the philosophy of mathematics, one of Wittgenstein's most central critical targets. This criticism leads him to interrogate the "superstitious dread and veneration by mathematicians in the face of a contradiction,"[21] as well as the whole conception of the work of the researcher in mathematics or mathematical logic that follows from the attempt to detect or preclude "hidden" contradictions. In particular, as Wittgenstein suggests, there is in fact no way that a "hidden contradiction" can vitiate a calculus as it is actually used. For if the contradiction remains "hidden," it has no effect on our actual practice of calculation; and if it is "discovered," then we need not act on it, and so again it can cause no harm. Thus:

> One may say, "From a contradiction everything would follow." The reply to that is: Well then, don't draw any conclusions from a contradiction; make that a rule. You might put it: There is always time to deal with a contradiction when we get to it. When we get to it, shouldn't we simply say, "This is no use—and we won't draw any conclusions from it"? (Diamond 1976, Lecture XXI, p. 209)

Elsewhere, Wittgenstein likens the situation of being faced with a contradiction to that of being given two conflicting orders, or being faced with two arrows pointing in opposite directions. That, in these situations, we do not know what the rules or orders are telling us to do and so, in that sense, "could" do *anything* "in accordance" with them, does not mean that we *must* do anything at all; we might simply abstain from acting. Or we might indeed take it that "anything is now permitted," but this would amount not so much to showing that the original calculus was out of order, as to giving up on the possibility of using a regular calculus to determine our action at all. Thus, there is no special need to worry about the presence of "hidden contradictions" and seek to develop a calculus.

This emphatically does not mean that we do not or even *should* not attempt to reason in accordance with the law of non-contradiction; indeed, Wittgenstein takes the fact that we do in fact do so, and regularly criticize those who violate it, to be an important and deep constitutive fact about (what we call) reasoning itself, such that anyone who did not reason in general in accordance with the law of noncontradiction, or respect its status as an overarching principle, would not be doing anything that we could recognize as reasoning or calculating at all. Also, we *can* and *do* construct our calculi with a view to avoiding – as much as is possible, anyway – the likelihood of encountering the situation of contradiction in which we, "entangled in our rules" as it were, are stopped and do not know how to go on. Wittgenstein's consideration of contradictions and their "status in civil life" does not show, therefore, that it is not an important and even *constitutive* element of our ordinary practices that we are committed, in practice, to avoiding them.[22] But it suffices to show that the existence of contradictions alone is not enough to completely vitiate these ordinary practices or render them ineffective.

This position about the role of contradiction is not at all implausible when applied to reasoning in a natural language such as English; for it is *extremely* plausible that there are contradictions in English, but this clearly does not make it rationally possible to "draw any conclusion at all" or vitiate the usefulness of reasoning in English. Nevertheless, when applied to "artificial" systems and techniques of calculation that we create, it is sufficiently counter-intuitive, or at least at odds with the ordinary particular self-conception of mathematicians and logicians, that it is regularly rejected by them as absurd or obviously incoherent. From the perspective of foundationalist assumptions, indeed, it can seem just obvious that the presence of a contradiction within a system, hidden or not, *must* have profoundly destructive consequences for the integrity of that system and cannot simply be handled in the offhand way that Wittgenstein suggests. This conception is regularly accompanied by a conception of logical systems or calculi as more or less "accurate" to the extra-logical reality that they concern, a conception which suggests that there could be calculi that are more or less effective and that a calculus containing contradictions would not be effective at all.

In the 1939 *Lectures*, Turing himself suggests that at least one of the problems with tolerating contradictions in a calculus is that the presence of a hidden contradiction in a calculus used for a technical application, say building a bridge, could lead to errors which cause the bridge to fall down.[23] Wittgenstein seizes on this claim and attempts over the course of the next several lectures to demonstrate that it is mistaken. That is, there are, according to Wittgenstein, *only* two ways in which our use of the calculus can lead to the bridge falling down: either because we use a wrong *physical* law (or get the value of a coefficient wrong, etc.) or because somebody makes a mistake in *calculation* and gets a wrong answer (although what counts as a wrong answer as opposed to a right one still must be somehow determined).[24] In either case, however, it is not a contradiction in the *calculus* that leads to the bridge falling down, and if such a contradiction actually arises we can choose to act on it as we like, or not at all. In any case, the technical *efficacy* and utility of the calculus is not adversely affected by the presence of a contradiction, and so there is no need for "foundational research" directed to assuring the universal absence of contradictions in our logical systems.

III

What, then, are the implications of Wittgenstein's way of looking at the significance of the results of Gödel and Turing for the issues of computation, human capacities, and finitude?

Interpretations of Gödel's theorem have spawned a large literature on issues of computationalism and the nature and capacities of the human mind. Much of this literature simply assumes Gödel's way of looking at his own result in terms of incompleteness, but Wittgenstein's way of looking at it evidently suggests an alternative. In particular, Gödel himself thought that the existence of the sentence *P* shows, for each formal system such as *Principia Mathematica,* the existence of a mathematical "truth" that that particular system cannot prove or disprove. Such "truths" are, according to Gödel's informal argument, accessible to the human mind in a way that essentially transcends the powers of any formal system; thus Gödel himself thought (e.g., van Atten 2006, p. 256) that his result demonstrated a *superlative* capacity of the human mind to grasp mathematical truths in excess of the powers of any formal system.[25] Subsequently, Lucas (1961) and Penrose (1994) have generalized this suggestion, holding that the combination of Gödel's and Turing's result show that the human mind (for instance that of a human mathematician) is not mechanical in the sense that it cannot be modeled by any formal system or Turing machine. Thus, both conclude, the human mind has capacities for grasping mathematical truths that exceed those of any machine or wholly mechanical system. There are certainly many problems with this argument, some of which have been pointed out over the years; in the present context, however, it is sufficient to note that Wittgenstein's different way of looking at the upshot of Gödel's result in fact provides a dramatic alternative to it.[26] As we saw, Wittgenstein's remarks on Gödel suggest that he take it as showing at least that the actual production of the Gödel sentence will lead to contradictions or antinomies, although it is not evident that these contradictions must have the profound destructive *significance* that foundationalist assumptions about mathematics portray them as having. In any case, however, Gödel's proof on the interpretation Wittgenstein probably intended *does* show that there is an essential limitation to the ability of any formal language to model itself *completely and consistently*; this is why the Gödel sentence, which "encodes" the logic of proof (and hence, on Wittgenstein's reading, truth) for the system as a whole, leads to contradiction and antinomy.

In a little-discussed 1985 paper, Putnam considers the implications for computationalism of taking Gödel's result in just this kind of way.[27] As Putnam notes, projects in artificial intelligence and cognitive science have relied centrally on the distinction between (actual) performance and (ideal) competence. That is, according to a longstanding conception originating from Chomsky, at least part of the aim of such projects is to give a description of how the mind is "supposed" to work, how we would be thinking if we were *ideally* competent. Both Harman and Chomsky himself have suggested that such an idealized "competence" description is indeed a description "of correct thinking in the normative sense."[28] However, even if there is such a description, would it be possible for us to know it? By way of a proof whose core is Gödel's proof itself, Putnam shows that it would not. That is: "if there is a complete computational description of our own prescriptive competence – a description of the way our minds ought to work, where the ought is the ought of deductive logic or inductive logic – then we cannot come to believe that that description is correct when our minds are in fact working according to the description" (p. 144). The reason for this is just the same as that underlying Gödel's and Turing's results – that it is impossible, on pain of contradiction, for a formal system completely to model itself. It follows that, as Putnam argues, if the aim of cognitive science is indeed to give such an idealized description of our own competence, then cognitive science is essentially looking for something that we *cannot* find. In particular, even if we did find what is in fact the "correct" description of our ideal competence, we could not know that it *was* the correct one.

What more general conclusion should we draw from this? As Putnam suggests, we may take the upshot in either an "optimistic" or a "pessimistic" way: "Like everything else, this theorem can be viewed either optimistically or pessimistically. The optimistic interpretation is: Isn't it wonderful! We always have the power to go beyond any reasoning that we can survey and see to be sound. Reflexive reflection cannot totally survey itself. The pessimistic interpretation is: How sad!" (p. 144). Here, the "optimistic" and "pessimistic" ways of looking at the failure of reflexive reason to survey itself essentially correspond to the two ways of looking at Gödel's result that we have already considered: Gödel's own, on one hand, and Wittgenstein's, on the other. In particular, the proponent of a Gödel-style interpretation sees the necessary failure of formal reason to survey itself as the sign of a superlative power or capacity, an ability of the human mind to non-formally exceed or "go beyond" all that formalism can model in itself. The proponent of the "pessimistic," Wittgenstein-style interpretation, on the other hand, takes the result wholly negatively – simply as showing that, as Putnam says, "reflexive reflection cannot totally survey itself," without taking this to imply any superlative capacity of the human mind.

With respect to computability, the analogue is apparently to take Turing's result itself wholly negatively – that is, as showing that it is not possible, on pain of contradiction (or at least paradox) for our rational procedures to model themselves completely. This suggests that there will be, among these, some infinitary procedures that, although perfectly determinate, are not effectively computable. This by itself does not suffice to show *what* these procedures actually are, or to guarantee our access to them. But such infinitary techniques, fixtures of human life that are not fixed, in their totality, by any finite symbolism, may be just what Wittgenstein is alluding to when, resolving the rule-following paradox of the *Philosophical Investigations,* he suggests that:

> 201. There is a way of grasping a rule which is *not* an *interpretation,* but which is shown in what we call 'obeying the rule' and 'going against it' from case to case.

And:

> 199. To understand a language means to be master of a technique.

Here, what Wittgenstein is suggesting is, importantly, *not* a superlative capacity of human thought to grasp "truths" or follow "procedures" that are inherently beyond the grasp of any mechanical system. Indeed, one of the central aims of the "rule-following considerations" is obviously to criticize any such conception of human ability to "leap beyond" all the finite examples and see an infinitary structure "all at once," a conception that yields metaphors of the use of a word being present all at once "in a queer way" and of grasping the entirety of a use of a word "in a flash."[29] This conception is also of a piece with the conception of rules as "rails laid to infinity" and thus as capable, by themselves, of determining infinite usage mechanically and completely.[30] On any of these metaphorical pictures, the whole use of a word – or the whole (infinite) extension of a mathematical series – is something that can be present "all at once" in the symbolism that expresses the rule. However, given any such symbolism, it is of course always possible to interpret it in various different ways. This is what leads to the problem to which section 201 gives an answer, the problem that "no course of action could be determined by a rule, because any course of action can be made out to accord with the rule." Given this, it looks as if it is indeed necessary to "give one interpretation after another," interpreting each (symbolic expression of a) rule with another until we realize that the second does no better than the first at guaranteeing the correct application, and so forth.

This paradox is more or less unavoidable, on the assumption that a rule as symbolically expressed must be able to determine its own infinite extension completely. This means that to resolve the paradox, we must get beyond the conception of rules according to which they are "self-interpreting," or capable of determining their own applications completely and without contradiction. To see that there is indeed a close analogy with Turing's own formal result here, consider again the details of Turing's formal argument for the undecidability of the halting problem. To establish this result, we posited (for *reductio*) a universal Turing machine capable of solving the general halting problem, and then considered whether it halts when given its own machine number. The result was the contradiction that it both does and does not halt, and accordingly that there can be no such machine (on pain of contradiction, at least). The general reason for this contradiction was that, in determining whether the machine with each description number halts, the posited machine must consider itself, and thus is apparently involved in an infinite regress. This regress is similar to the regress of symbolic interpretations that occurs inevitably if we assume that a rule must be able to determine its *own* application. In particular, the demand that the rule determine the application of another one is essentially similar to the requirement that a particular Turing machine determine the halting status of another one; and the demand that a rule must be able ultimately to determine its own application is then analogous to the requirement that a universal Turing machine determine its own halting status. In both cases, the demand of self-determination leads to an intractable paradox that shows that this demand is not completely and consistently satisfiable. Just as there is no mechanical procedure that solves the general halting problem, and thus no machine that can ultimately guarantee whether it, itself, halts, there is thus no way for any rule to *guarantee* the correctness of its own infinite application.

This does not mean, of course, that the correctness of a rule's application *is* guaranteed by something else, for instance (as we may now be tempted to think) an ineffable insight, or a power of human judgment or discernment that "essentially exceeds" anything mechanical. Rather, as Wittgenstein repeatedly emphasizes, the right move when faced with the gap between the demand that the correct application *must* be determined by something "present to mind" and the rule's incapacity to do so is not to appeal to any supplemental figure of ineffable force to fill the gap but rather to relax the demand that produces it. The result is that there is indeed no finite, symbolic expression – and hence nothing that can be "present to mind" all at once, or already implicit in any determination of first principles or fundamental axioms – that indeed suffices *by itself and outside of its practical context* to determine and guarantee the distinction between correct and incorrect application in all cases. On the other hand, a “technique” or “practice” is, rather, essentially something that unfolds over time, and in the relationship between people.[31] Thus, no symbolic expression or finitely capturable “capacity” has, as we may now say, the absolute *force* of a law which would be capable of determining the distinction between correctness and incorrectness all by itself. The task of philosophical criticism and “therapy” then shifts to a radical diagnosis and replacement of the assumption of (and the demand for) such a force.

IV

I have suggested, then, that Wittgenstein’s way of looking at the results of Gödel and, implicitly, Turing, gives us a way of conceiving of their implications that, although it bears important implications for the question of computationalism, does *not* tend to show (as the Gödel-Lucas-Penrose interpretation alleges) that the “mind is not a formal system” in that it has access to mathematical truths in excess of the grasp of any formal system. This does not, of course, imply that Wittgenstein would have agreed to the opposite claim that the “mind is a formal system,” or even would take it, ultimately, to have much of a clear sense.[32] Indeed, one of the deepest aims of the whole line of argument that is developed in the rule-following considerations is to formulate a kind of critical resistance to pictures that identify human techniques and capacities with what are conceived of as the capacities of formal systems and as wholly present in their underlying structure. This resistance is deeply connected to Wittgenstein’s critical interrogation of the conception of “logical inexorability” and “necessity” that these pictures suggest, and ultimately to his more basic inquiry into the sources of “logical necessity” and “rational compulsion” themselves.[33] According to Wittgenstein, when we picture to ourselves the compulsory force of logical rules, we are led to think of logic as a kind of machinery underlying our actual practices of reasoning and inferring. Such a machinery would determine “in advance” and without exception the correct practices of logical inference and derivation; in this respect, it is akin to a kind of “super-rigid” machine that contains all of its possible actions in itself by virtue of its ideal construction:

> A machine as symbolizing its action: the action of a machine—I might say at first – seems to be there in it from the start. What does that mean?—If we know the machine, everything else, that is its movement, seems to be already completely determined. (*PI* 193, p. 66)

If we are to conceive of an actual machine this way, we must of course forget or abstract from the empirical possibility of its parts “bending, breaking off, melting, and so on.” It is in fact just such an abstraction that is essential to our symbolizing the machine as such (for instance by means of a functional blueprint) and it is this alone that permits the movement of formalization whereby we consider *any* actually existing machine actually to “realize” or “amount to” an “ideal machine” such as a Turing machine or a computer. Wittgenstein’s point is not that this forgetting or idealization is not sometimes justified, but that it encourages a conception of logical necessity that is itself deeply misleading. For:

> …when we reflect that a machine could also have moved differently it may look as if the way it moves must be contained in the machine-as-symbol far more determinately than in the actual machine. As if it were not enough for the movements in question to be empirically determined in advance, but they had to be really – in a mysterious sense – already present. And it is quite true: the movement of the machine-as-symbol is predetermined in a different sense from that in which the movement of any given actual machine is predetermined. (*PI* 193, p. 66)

The ideology of the super-rigid machine, in which all of its movements are already present, is thus the same as that of rules as self-interpreting or as rails laid to infinity: in both cases, it is only by virtue of a movement of idealization and abstraction from actual cases that we gain the conception of an underlying presence that is actually *effectively* capable of determining the entirety of an infinite extension in advance. And this conception of a presence as capable of such an infinite determination is itself the same as the conception of the force of logical determination as that of a “super-hard” or inexorable law.

How, though, do we first arrive at such a conception? In a passage from the 1939 lectures devoted to the idea of a super-rigid "logical machinery," Wittgenstein compares the source of the underlying idea of the "inexorability" of logical law to that of the inexorability of the law as such:

> Perhaps it would help to take the example of a perfectly inexorable or infinitely hard law, which condemns a man to death.
> A certain society condemns a man to death for a crime. But then a time comes when some judges condemn every person who has done so-and-so, but others let some go. One can then speak of an inexorable judge or a lenient judge. In a similar way, one may speak of an inexorable law or a lenient law, meaning that it fixes the penalty absolutely or has loopholes. But one can also speak of an inexorable law in another sense. One may say that the law condemns him to death, whether or not the judges do so. And so one says that, even though the judge may be lenient, the law is always inexorable. Thus we have the idea of a kind of super-hardness.
> How does the picture come into our minds? We first draw a parallel in the expressions used in speaking of the judge and in speaking of the law: we say "the judge condemns him" and also "the law condemns him". We then say of the law that it is inexorable – and then it seems as though the law were more inexorable than any judge – you cannot even imagine that the law should be lenient. (Diamond 1976, p. 197)

The image of the ideal inexorability of the law is thus, like the picture of the super-rigid machine itself, produced out of a kind of false parallel or crossing between two expressions, one that is used ordinarily to describe judges, and another that is used (perhaps metaphorically) to speak of the law itself. The diagnosis does not imply that either the picture of the inexorability of the law or that of the super-rigidity of logical machinery is completely out of order, false, or nonsensical; either picture may indeed have its legitimate uses. However, it does suffice to show that both are grounded in a kind of essential confusion:

> …if I say that there is no such thing as the super-rigidity of logic, the real point is to explain where this idea of

> super-rigidity comes from – to show that the idea of *super-rigidity* does not come from the same source which the idea of *rigidity* comes from. The idea of rigidity comes from comparing things like butter and elastic with things like iron and steel. But the idea of super-rigidity comes from the interference of two pictures – like the idea of the super-inexorability of the law. First we have: "The law condemns", "The judge condemns". Then we are led by the parallel use of the pictures to a point where we are inclined to use a superlative. We have then to show the sources of this superlative, and that it doesn't come from the source the ordinary idea comes from. (Diamond 1976, p. 199)

That is, it is only by means of this sort of crossing or confusion between two pictures that we gain the idea of the inexorable law of logic, and hence of its *force* in regulating our life and practices. This corresponds, as we have seen, to what is often called "normative" force, and conceived as a distinctive *kind* of force that is both distinct from and stronger and more inexorable than any kind of empirical or physical force. This conception is the same as that of the normativity of logical rules, or of the "self-applying" rule which is able to determine its own application in a logical and normative sense. The diagnosis of the *origin* of these pictures in a crossing or confusion between the empirical attributes of actual machines (for which it makes sense to say that one is more or less rigid than another) and the posited non-empirical attributes of an ideal machine makes it clear that the conception of the inexorable normative force of logic is itself grounded in such a confusion, and dissipates with its successful diagnosis.

This does not mean that the "normativity" that is involved in ordinary practices of rational deliberation and calculation which proceed, in part, by way of the citation and discussion of explicit rules, should be dismissed as simply illusory or fictitious. But it does imply that the underlying source of this normativity cannot be in these rules themselves (as symbolically expressed or expressible), but must have a deeper ground in the kinds of "agreement" and "attunement" that constitute our "forms of life," and thereby precondition the possibility of all techniques and practices as such. We can draw much the same conclusion, moreover, about the "normativity" exhibited by instruments and techniques of calculation, including actual symbolic computing machines or computers. Here, as the results of Turing and Wittgenstein both show, the ability of such instruments and devices to determine the distinction between correct and incorrect results (or, for instance, correct and incorrect ways of extending a function) does not and cannot rest entirely on anything given or wholly determined by the actual construction of the instruments and devices themselves. It depends, instead, on the pre-existing practices, techniques, and ways of life in which these instruments and devices have their normal roles.[34]

If this is right, though, and the practical interpretation of a given object or piece of machinery depends on our pre-existing ability to distinguish between "correct" and "incorrect" instances of computation, then the normativity involved in this distinction again essentially *cannot* be given wholly by any computable system of rules itself. It must, instead, already be a precondition of our ability to take any set of symbols or physical system *as* the expression or implementation of any such system. This implies, again, that any attempt to ground our judgments of correctness or incorrectness in the inexorable force of a "logical machine" that determines the interpretation of its own symbolism, or the implementation of its own computations, all by itself and outside the context of any human practices, must inevitably fail. Our practical attitudes toward the rules embodied in actual computers, and the kinds of normative force they represent or enforce, are to the contrary very much aspects of our everyday lives and practices with them, and accordingly cannot be separated from these ordinary practices. The normativity that we expect from, and regularly find, in the actions of computers is not simply an outcome of their actual construction or their "ideal" architecture, but is rather possible only on the basis of the kinds of "agreement" that first enable us to engage in shared practices at all. As Wittgenstein emphasizes, this agreement is not underlain or guaranteed by any technical or technological form of regularity or repetition. At the same time, however, it is not at all a *contingent* agreement on specific (as it may be, "historically situated") practices, norms, or conventions. For:

> …the logical 'must' is a component part of the propositions of logic, and these are not propositions of human natural history. If what a proposition of logic said was: Human beings agree with one another in such and such ways (and that would be the form of the natural-historical proposition), then its contradictory would say that

> there is here a *lack* of agreement. Not, that there is an agreement of another kind. (*RFM* VI-49, p. 353)[35]

NOTES

1. Some discussions among Wittgenstein, Turing, and Wittgenstein's student Alister Watson had reportedly taken place earlier, in the summer of 1937, but there is no record of these. (Hodes 1983, pp. 109, 136 – cited in Floyd and Putnam (2000))

2. Wittgenstein (1939), p. 13.

3. Wittgenstein (1939), p. 14.

4. "We have said that the computable numbers are those whose decimals are calculable by finite means. This requires rather more explicit definition … For the present I shall only say that the justification lies in the fact that the human memory is necessarily limited" (p. 59); "The behaviour of the computer at any moment is determined by the symbols which he is observing, and his 'state of mind' at that moment. We may suppose that there is a bound B to the number of symbols or squares which the computer can observe at one moment … We will also suppose that the number of states of mind which need to be taken into account is finite. The reasons for this are of the same character as those which restrict the number of symbols" (pp. 75–76). Turing also emphasizes (p. 79) that there must at any moment be a symbolically stateable description which, if the computer broke off work at any particular stage, would determinately instruct another as to how to continue.

5. More specifically, H combines the universal machine U with a "decision machine" D which, when given the description number of any particular machine, determines whether that machine halts.

6. The demonstration is on pp. 72–73.

7. Since, as Turing argues, "If the negation of what Gödel has shown had been proved … we should have an immediate solution to the Entscheidungsproblem," (p. 85) it follows that Turing's result – that there is no solution to the Entscheidungsproblem – implies Gödel's first incompleteness theorem.

8. It may be objected that the original proof of Cantor's theorem, which establishes the superiority of the cardinality of the power set over the initial set, does not make use of syntactic reasoning (I owe this objection to a discussion with John Bova, although it does not represent his view). But: i) since Cantor's proof does not obviously involve entertaining or carrying out any infinite procedure, it is not clear that it is an instance of the "diagonal procedure" at all; and ii) insofar as it involves the assumption for reduction of a 1-1 *correspondence* between sets and their subsets, it does involve (at least where the sets are infinite) something like a comparison of the infinite set with its finite elements, something very similar to the "comparison" of an infinitary procedure with its finitely expressed rule in syntactical diagonalization.

[9] *RFM* II – 38: “Here it is important to grasp the relationship between a series in the non-mathematical sense and one in the mathematical sense. It is of course clear that in mathematics we do not use the word ‘series of numbers’ in the sense ‘series of numerical signs,’ even though, of course, there is also a connexion between the use of the one expression and of the other. … A ‘series’ in the mathematical sense is a method of construction for series of linguistic expressions” (p. 136).

[10] Gödel’s original article is titled “On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I” and mentions “incompleteness” only once, in a footnote: “The true reason for the incompleteness which attaches to all formal systems of mathematics lies – as will be shown in Part II of this paper – in the fact that the formation of higher and higher types can be continued into the transfinite (cf. D. Hilbert 'Über das Unendliche', *Math. Ann.* 95, p. 184), while, in every formal system, only countably many are available. Namely, one can show that the undecidable sentences which have been constructed here always become decidable through adjunction of sufficiently high types (e.g. of the type w to the system **P**). A similar result holds for the axiom systems of set theory.” (footnote 48a, pp. 28–29).

[11] Thus: “From the remark that [R(q); q] asserts its own unprovability it follows immediately that [R(q); q] is true, since [R(q); q] is indeed unprovable (because it is undecidable). The proposition undecidable in the system *PM* is thus decided by metamathematical arguments” (p. 9). As Gödel emphasizes, this remark comes as part of a “sketch of the main ideas of the proof” that does not make “any claim to rigor” (p. 6).

[12] Floyd and Putnam (2000), p. 625. For further discussion, see Steiner (2001), Bays (2004), and Floyd and Putnam (2006).

[13] Floyd and Putnam (2000), pp. 624–26. A system is ω-inconsistent if, for some property T of natural numbers formulable in the system, the system proves T(0), T(1), and so forth, but also proves *that there is some* natural number *n* such that ~T(*n*). Note that a system may be ω-inconsistent but still consistent.

[14] Floyd and Putnam (2000), p. 632.

[15] Cf. *PI* 136.

[16] Cf. Priest (2004).

[17] Priest (2004), p. 213: “Consider the sentence *A*, of the form ‘<*A*> is not provable’ – this sentence is not provable – angle brackets represent some naming device. Here, provability is to be understood in the naïve sense of being demonstrated by some argument or other. If *A* is provable, then, since what is provable is true, *A* is true; so <*A*> is not provable. Hence, <*A*> is not provable. But we have just proved this; that is, <A> is provable. This is a version of the ‘Knower paradox’. Sometimes it is called ‘Gödel’s paradox’. In fact, if one identifies truth with provability, as does Wittgenstein, Gödel’s paradox and the liar collapse into each other.”

[18] Cf. Priest (2004), p. 223: “According to the model-theoretic account of truth, the equivalence (I) [viz. the *interpretation* of the Gödel sentence as saying ‘*P* iff *P* is not provable in *Principia*’] is unproblematic. In the context that

Wittgenstein is operating in, it is not, and this allows him to question it. In particular, he can ask exactly what the right-hand side means. This allows him to take the discussion into areas beyond those normally countenanced in discussions of Gödel's theorem. In particular, Wittgenstein deploys the idea that the meaning of a sentence is determined by its proof conditions. In virtue of the fact that there are object-level proofs and meta-level proofs (to put it in modern terminology), this still leaves the notions concerned in (I) ambiguous. Except for one circumstance [i.e. that *Principia* is inconsistent], however, he thinks that once one clarifies the relevant meanings, the equivalence (I) should be rejected. In this case, no contradiction is forthcoming."

19. Of course, *PM* (etc.) contain rules establishing that "from a contradiction, anything follows." However, it is clear that there are grounds for being skeptical that, even if this is true, and "anything follows" in the formal sense, a single contradiction is indeed enough to render the calculus useless; see below.

20. What, though, *is* in fact the ultimate basis for our belief in the consistency of a system such as *Principia Mathematica*? The usual basis is model-theoretical arguments, but if we dispute that a model-theoretical notion of truth is appropriate here, we may well doubt these arguments. Moore (2001, p. 177–180) argues in a related context that we can take the set-theoretical axiom framework ZF to be sound (and hence consistent) if we can "recognize" its axioms as intuitively correct. However, what is the ground for such a "recognition"? In any case, because of the Gödel sentence, Principa Mathematica is certainly *not* consistent if we allow the "disquotational" rule Prov(x) -> x.

21. *RFM*, I, appendix III, remark 17 (p. 122).

22. *PI* 125.

23. Lecture 21, p. 211.

24. Lecture 22, p. 211.

25. Thus, Gödel wrote in 1963: "Before my results had been obtained it was conjectured that any precisely formulated mathematical yes or no question can be decided by the mechanical rules of logical inference on the basis of a few mathematical axioms. In 1931 I proved that this is not so. i.e.: No matter what & how many axioms are chosen there always exist number theoretical yes or no questions which cannot be decided from these axioms. Combining the proof of this result with Turing's theory of computing machines one arrives at the following conclusion: Either there exist infinitely many number theoretical questions which the human mind is unable to answer or the human mind … contains an element totally different from a finite combinatorial mechanism, such as a nerve net acting like an electronic computer. I hope I shall be able to prove on mathematical, philosophical, & psychological grounds that the second alternative … holds" (van Atten, 2006).

26. The most important of the problems with the Lucas-Penrose argument in the present context is that it requires the actual *manifestation* of a true formula T of arithmetic such that a certain actual computer C *cannot* give a proof of T, but there is a human mind, M, that can. No one has ever actually manifested such a formula, and there is reason to think that it indeed *cannot* be demonstrated by any

effective procedure. To see this, observe that to find such a formula, given any *actual* computer C, we would have to first distinguish those portions of its output that actually count as proofs from those that do not. As Priest (1994, pp. 111–115) argues, there is, however, probably no effective way to do this; and even if we do, the set of theorems T proved by any actual computer C will probably turn out to be inconsistent. Priest writes: "The only way … that offers any hope of getting T to be consistent is to suppose that M (and so any C which is supposed to be M) is not only a mathematical mind but an ideal mathematical mind, that never makes mistakes of any kind: either of memory, inference, judgment, or output. But this is sufficient to destroy the argument. After all, the only candidate for a mind of this kind is God's. So at best, we have a (theo)logical proof that God is not a computer" (p. 113).

[27] Putnam (1985).

[28] Putnam (1985), p. 149.

[29] *PI* 191, 195.

[30] *PI* 218.

[31] Cf. *PI* 199: "Is what we call 'obeying a rule' something that it would be possible for only *one* man to do, and to do only *once* in his life?—This is of course a note on the grammar of the expression 'to obey a rule'.
It is not possible that there should have been only one occasion on which only one person obeyed a rule. It is not possible that there should have been only one occasion on which a report was made, an order given or understood; and so on.—To obey a rule, to make a report, to give an order, to play a game of chess, are *customs* (uses, institutions).
To understand a sentence means to understand a language. To understand a language means to be master of a technique."

[32] Cf. *PI* 359–60.

[33] Cf. *RFM* I–117: "In what sense is logical argument a compulsion?—'After all you grant this and this; so you must also grant this!' That is the way of compelling someone. That is to say, one can in fact compel people to admit something in this way.—Just as one can e.g. compel someone to go over there by pointing over there with a bidding gesture of the hand" (p. 81).

[34] This is particularly evident in connection with what have been called "triviality" arguments about computation and effectiveness. According to such arguments, every physical object (or every object of a certain, very minimally specified level of operational complexity) at every time trivially implements every possible computation, since there is always *some* function that maps the internal physical states of the object onto the computational states involved in carrying out any particular computation. Thus, on Searle's memorable formulation: "…the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar, if it is a big enough wall it is implementing any program, including any program implemented in the brain" (Searle 1992). Such arguments have been used as well, most notably by

Putnam himself (in Putnam 1991) to argue against computational and functionalist theories of mind by showing that there is no *unique* functional description that characterizes the operation of the human brain (or any other mechanical system) at any time, and accordingly no hope for the "computationalist" project which attempts to discover such a (unique) description. As respondents to the triviality arguments (e.g. Block 1995, Chalmers 1995) have pointed out, we can solve the problem of triviality if – and only if – we can already *presuppose* a distinction between the correct and incorrect functioning of the machine. Thus, for instance, I can "interpret" the machinery in front of me as calculating the "plus" function only if I am in a position to distinguish between correct and incorrect responses to (what *I* interpret as) a query, for instance "2+3=?"; and I will indeed be *inclined* to interpret the machinery as calculating the "plus" function only if I can assume that it *reliably* gives (what I deem to be) correct responses to this query. It is important to note, however, that there is nothing that guarantees such reliability (any actual machine might "break down" at any moment), and no sharp line between what kinds of behavior count as "reliable" and what evidences "unreliability" in this sense.

35. Earlier and shorter versions of this paper were presented at the Austrian Ludwig Wittgenstein Symposium in Kirchberg am Wechsel, August, 2009, and at the North American Wittgenstein Society in San Francisco, March 2010. I wish to thank John Bova for extensive discussions about the issues considered here, and Jack Woods for his extensive and helpful commentary on the paper at the NAWS session.

BIBLIOGRAPHY

Bays, T. (2004), "On Floyd and Putnam on Wittgenstein on Gödel," *The Journal of Philosophy* 101(4): 197–210.

Block, N. (1995), "The Mind as the Software of the Brain," in Smith, D. and Osherson, E. (eds.), *Invitation to Cognitive Science: Thinking*. 2nd ed. Cambridge, MA: MIT Press.

Chalmers, D. (1995), "On Implementing a Computation", *Minds and Machines* 4: 391–402.

Diamond, C. (ed.) (1976), *Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge, 1939*. Chicago: University of Chicago Press.

Floyd, J. and Putnam, H. (2000), "A Note on Wittgenstein's 'Notorious Paragraph' about the Gödel Theorem", *Journal of Philosophy* 97(11): 624–632.

Floyd, J. and Putnam, H. (2006), "Bays, Steiner and Wittgenstein's "Notorious" Paragraph about the Gödel theorem," *The Journal of Philosophy* 103: 101–110.

Gödel, K. (1931), “On Formally Undecidable Propositions of Principia Mathematica and Related Systems I”, translated and reprinted in Davis, M. (ed.) (2004), *The Undecidable*. Mineola: Dover.

Lucas, J.R. (1961), “Minds, Machines, and Gödel,” *Philosophy* 36: 112–127.

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

Priest, G. (1994), “Gödel’s Theorem and Creativity,” in Dartnall, T. (ed.), *Artificial Intelligence and Creativity*. Dordrecht: Kluwer.

Priest, G. (2002), *Beyond the Limits of Thought*. 2nd ed. Oxford: Clarendon.

Priest, G. (2004), “Wittgenstein's Remarks on Gödel's Theorem,” in Kölbel, M. and Weiss, B. (eds.), *Wittgenstein’s Lasting Significance.* London: Routledge.

Putnam, H. (1985), “Reflexive Reflections”, *Erkenntnis* 22(1/3): 143–153.

Putnam, H. (1991), *Representation and Reality*. Cambridge, MA: MIT Press.

Steiner, M. (2001), “Wittgenstein as his Own Worst Enemy: The Case of Gödel’s Theorem,” *Philosophia Mathematica* 9: 257–279.

Turing, A. (1936), “On Computable Numbers, with an Application to the Entscheidungsproblem,” in Copeland, B.J. (ed.), *The Essential Turing*. Oxford: Clarendon, 58–93.

Van Atten, M. (2006), “Two Draft Letters from Gödel on Self-Knowledge of Reason,” *Philosophia Mathematica* III(14): 255–261.

Wittgenstein, L. (1978), *Remarks on the Foundations of Mathematics*. Rev. ed. Cambridge, MA: MIT Press.

Wittgenstein, L. (1979), *Notebooks 1914–1916*. 2nd ed. Chicago: University of Chicago Press.

Wittgenstein, L. (2001), *Philosophical Investigations*. 3rd ed. Oxford: Blackwell.