

The Mathematical Objection: Turing, Gödel, and Penrose on the Mind

Jack Copeland, July 2008

Turing in 1947: "There can be no machine which will distinguish provable formulae of the system from unprovable ... On the other hand if a mathematician is confronted with such a problem he would search around and find new methods of proof."

A central aspect of mathematical creativity is the dreaming up of new methods—new methods for proving theorems, solving problems, and so forth. Is the creative process in human mathematicians always computable (assuming an unbounded supply of mathematicians, paper, ink, time)? There are various ways of expressing this question more precisely. One is: is the sequence of new methods ... m_i, m_j, \dots produced over time by the idealised mathematical community a computable sequence in the sense of Turing 1936?

I certainly shan't answer this question here, and my approach will be largely historical and textual, but I hope to draw attention to some interesting features of the landscape.

Turing on the Inexhaustibility of New Methods

The idea that the devising of new methods may be a non-mechanical aspect of mathematics is present in Turing's logical work from an early stage.

Turing in 1936

"Let δ be a sequence [e.g. 10111001...] whose n -th figure is 1 or 0 according as n is or is not satisfactory. It is an immediate consequence of the theorem of § 8 that δ is not computable. It is (so far as we know at present) possible that any assigned number of figures of δ can be calculated, but not by a uniform process. When sufficiently many figures of δ have been calculated, an essentially new method is necessary in order to obtain more figures." (Turing in 'On Computable Numbers', 1936)

This is a very interesting quotation. Turing is envisaging the possibility that the human mathematician can calculate any desired number of digits of an *uncomputable* sequence by virtue of creating new methods when necessary. Gualtiero Piccinini discusses this quotation in his 2003 paper 'Alan Turing and the Mathematical Objection'. Gualtiero also discusses a number of the other quotations from Turing that I discuss below. I hope our conclusions are in agreement—I'm sure he'll tell me if they aren't!

Turing during the Second World War

Letter from Turing to Max Newman, written at The Crown, Shenley Brook End, circa 1940:

"Intuition, Inspiration, Ingenuity

I am not sure whether my use of the word 'intuition' is right or whether your 'inspiration' would be better. ...

Turing goes on to discuss this a little and comes down in favour of his own term 'intuition'. Then he continues:

"The straightforward unsolvability or incompleteness results about systems of logic amount to this

- α) One cannot expect to be able to solve the Entscheidungsproblem for a system
- β) One cannot expect that a system will cover all possible methods of proof."

Turing is putting an interesting spin on the incompleteness results, which are usually stated in terms of there being true mathematical statements that are not provable. On Turing's way of looking at matters, the incompleteness results show that no single system of logic can include all methods of proof. He advocates a progression of logical systems, each more inclusive than its predecessors. He calls these ordinal logics. The letter continues:

"[W]e ... make proofs ... by hitting on one and then checking up to see that it is right. ... When one takes β) into account one has to admit that not one but many methods of checking up are needed. In writing about ordinal logics I had this kind of idea in mind."

Turing's Multi-Machine Picture of Mathematics

In a subsequent letter to Newman, Turing continues this discussion and sets out what I will call his multi-machine picture of mathematics. Gualtiero has also drawn attention to the multi-machine picture (although he does not use this term).

Letter from Turing to Newman, written at King's College, Cambridge, circa 1940:

"I think you take a much more radically Hilbertian attitude about mathematics than I do. You say 'If all this whole formal outfit is not about finding proofs which can be checked on a machine it's difficult to know what it is about'. [D]o you have in mind that there is (or should be or could be, but has not been actually described anywhere) some fixed machine ... and that the formal outfit is, as it were about this machine. If you take this attitude ... there is little more to be said: we simply have to get used to the technique of this machine and resign ourselves to the fact that there are some problems to which we can never get the answer. ... However I don't think you really hold quite this attitude because you admit that in the case of the Gödel example ... there is a fairly definite idea of a true formula which is quite different from the idea of a provable one. ...

If you think of various machines I don't see your difficulty. One imagines different machines allowing different sets of proofs, and by choosing a suitable machine one can approximate 'truth' by 'provability' better than with a less suitable machine, and can in a sense approximate it as well as you please. The choice of a ...

machine involves intuition, ... or as [an] alternative one may go straight for the proof and this again requires intuition."

So in this picture the role of intuition—or 'inspiration', or creativity in the sense under discussion—is localised very precisely. Intuition is responsible for the selection of the appropriate theorem-proving machine, the appropriate Turing machine, and the rest is mechanical. I shall return in later sections to Turing's multi-machine picture of mathematics.

Penrose and Post on Intuition and Creativity

"[H]uman intuition and insight ... cannot be reduced to any set of computational rules. ... Gödel's theorem indeed shows this, and provides the foundation of my argument that there must be more to human thinking than can ever be achieved by a computer." (Penrose, *Shadows of the Mind*: 65)

Although generally known as the 'Gödel argument', this form of objection was in fact anticipated by Emil Post as early as 1921. In a subsequent account of this anticipation Post wrote:

"The logical process is essentially creative. This conclusion ... makes of the mathematician much more than a kind of clever being who can do quickly what a *machine* could do ultimately. We see that a *machine* would never give a complete logic; for once the machine is made *we* could prove a theorem it does not prove." (Post in 1941)

Turing's Formulation of the 'Mathematical Objection'

Turing calls this line of argument the 'Mathematical Objection' to machine intelligence. He gave this succinct statement of the Mathematical Objection in his 1948 report to the National Physical Laboratory:

"Recently the theorem of Gödel and related results ... have shown that if one tries to use machines for such purposes as determining the truth or falsity of mathematical theorems and one is not willing to tolerate an occasional wrong result, then any given machine will in some cases be unable to give an answer at all. On the other hand the human intelligence seems to be able to find methods of ever-increasing power for dealing with such problems[,] 'transcending' the methods available to machines." (Turing, 'Intelligent Machinery': 410-11)

Turing states the argument with more detail in his 1950 article 'Computing Machinery and Intelligence':

"The questions that we know the machines must fail on are of this type, 'Consider the machine specified as follows . . . Will this machine ever answer "Yes" to any question?' The dots are to be replaced by a description of some machine in a standard form. When the machine described bears a certain comparatively simple relation to the machine which is under interrogation, it can be shown that the answer is either wrong or not forthcoming. This is the mathematical result: it is argued that it proves a disability of machines to which the human intellect is not subject." (Turing, 'Computing Machinery and Intelligence': 451)

So that's the objection that Turing is going to answer. Before looking at how his answer goes I want to talk a bit about the weakness as I see it of Penrose's notorious attempt to make use of the Mathematical Objection or 'Gödel argument'.

Penrose's Confession

Penrose makes an interesting admission in his book *Shadows of the Mind*. Casual readers of the book could be forgiven if they did not notice this admission, which is tucked away inconspicuously in the middle of a chapter entitled 'Quantum theory and the brain' (chapter 7). There Penrose says:

"[T]he arguments of Part I of this book can be applied equally well against an oracle-machine model of mathematical understanding as they were against the Turing-machine model, almost without change." (Penrose, *Shadows of the Mind*: 380)

An oracle machine is a Turing machine plus what Turing calls an 'oracle'—a black box able to produce the values of some function that is not computable by any Turing machine. So adding the box to the machine is a bit like adding a new, independent axiom to a logical system—as a result of the addition, the apparatus can prove more than it was previously capable of proving.

What Penrose describes as the *first-order o*-machines are those whose oracle returns the values of the Turing-machine halting function. So a first-order *o*-machine can say of each Turing machine whether or not it halts. The *second-order o*-machines are those with an oracle that can say whether or not any given first-order *o*-machine eventually halts if set in motion with such-and-such a number inscribed on its tape; and so on for third-order, and in general α -order.

Penrose's argument was originally marketed as demonstrating that human mathematicians are not Turing machines. But the argument appears to be so powerful that it can equally well be employed to show that, for every ordinal number α , human mathematicians are not α -order oracle machines. Penrose's argument moves relentlessly up through the orders, stopping nowhere. This discovery evidently disconcerts Penrose:

"The final conclusion of all this is rather alarming. For it suggests that we must seek a non-computable physical theory that reaches beyond every [recursive] level of oracle machines (and perhaps beyond). No doubt there are readers who believe that the last vestige of credibility of my argument has disappeared at this stage! I certainly should not blame any reader for feeling this way." (Penrose, *Shadows of the Mind*: 381)

That's a candid statement. Penrose does hint at a way out of his difficulty.

Penrose's Way Out

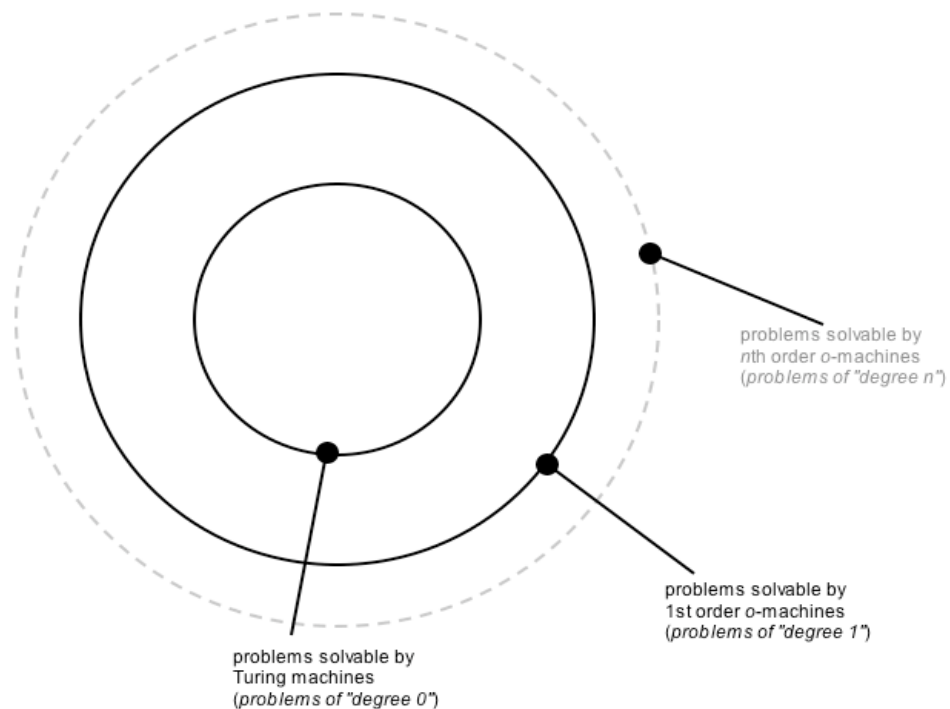
"[I]t need not be the case that human mathematical understanding is in principle as powerful as *any* oracle machine at all. ... [T]he conclusion **G** ["Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth"] does *not* necessarily imply that human insight is powerful

enough, in principle, to solve each instance of the halting problem. Thus, we need not necessarily conclude that the physical laws that we seek reach, in principle, beyond every computable level of oracle machine (or even reach the first order). We need only seek something that is not equivalent to *any* specific oracle machine (including also the *zeroth*-order machines, which are Turing machines). Physical laws could perhaps lead to something that is just *different*." (Penrose, *Shadows of the Mind*: 381)

On that enigmatic note Penrose leaves it. Just when we seem to be approaching a crucial part of the exposition, he suddenly falls silent.

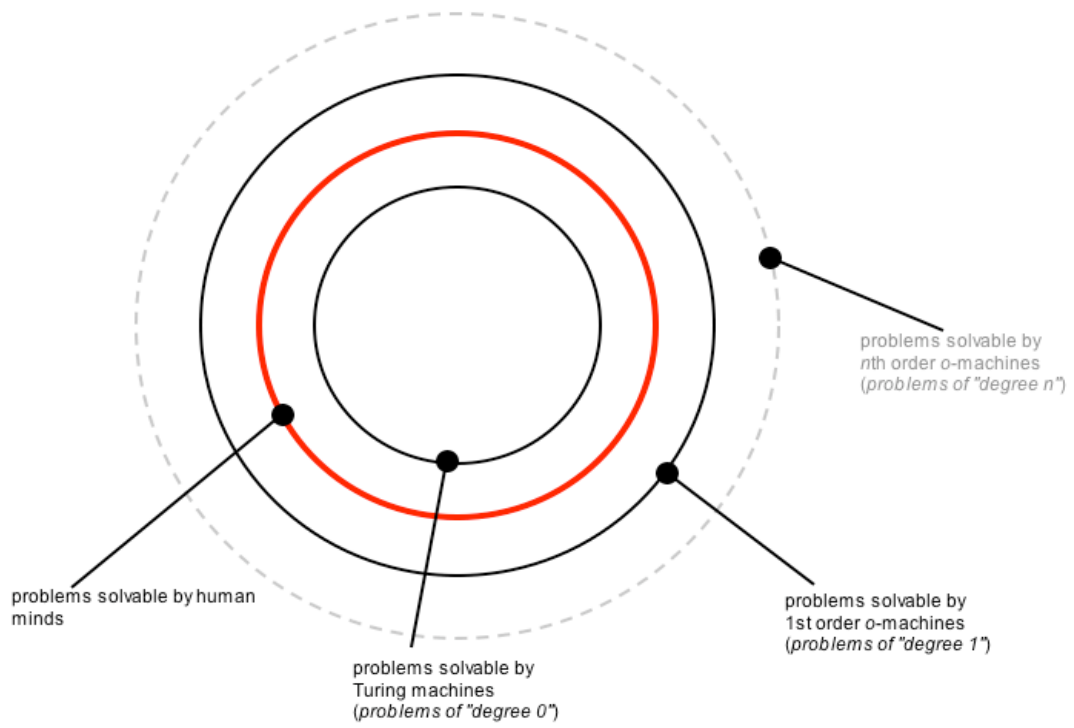
What did Penrose mean? Let me try to express Penrose's way out, and the difficulty involved in it, by means of a series of pictures.

The O-machines



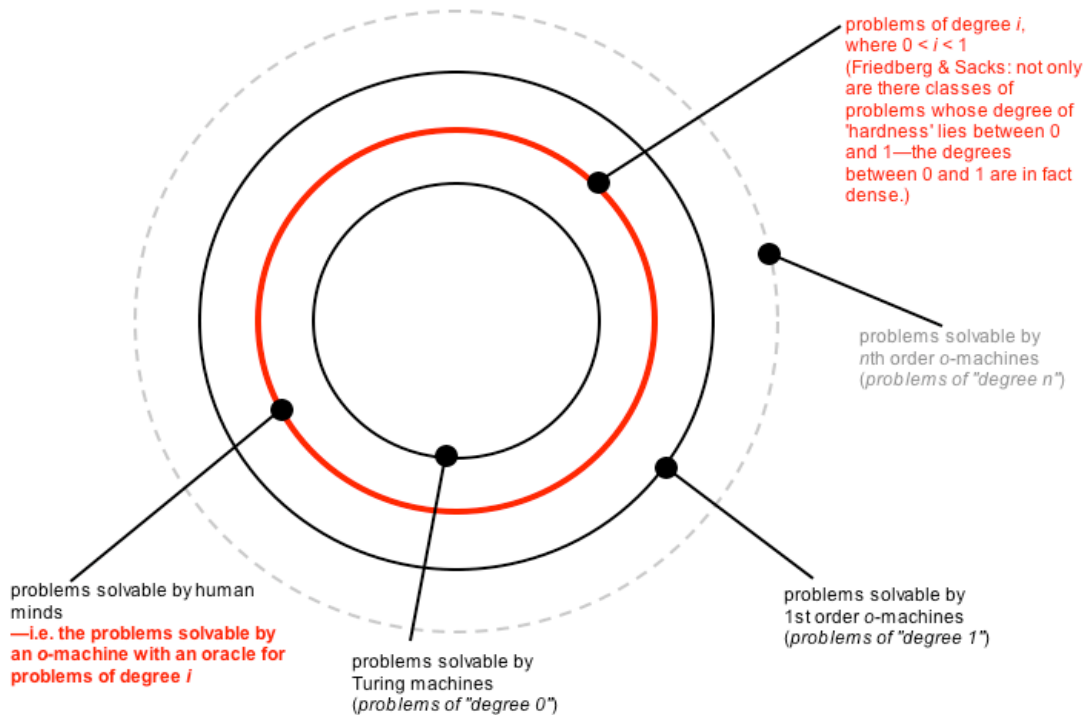
Penrose's Conjecture

Penrose said, 'we need not necessarily conclude that the physical laws that we seek reach, in principle, beyond every computable level of oracle machine (or even reach the first order). We need only seek something that is not equivalent to *any* specific oracle machine (including also the *zeroth*-order machines, which are Turing machines).' So perhaps the problems solvable by the human mind are positioned like this:



But...

But then it seems that the Gödel argument can simply be reapplied to show that human mathematicians are not *that* oracle machine either! For the new circle also corresponds to an oracle machine, as the next diagram shows. Once you open the door to the Gödel argument, it pursues you relentlessly. Wherever Penrose tries to draw the circle and say 'There! That's the mind!', the Gödel argument can be applied to show that no, that is not the mind.



Penrose's Last Word

In a later publication (1996) Penrose did briefly revisit the problem to which he drew attention in his 'confession'. John Lucas, an earlier exponent of the Gödel argument and no materialist, was happy to conclude from the argument that:

"[N]o scientific enquiry can ever exhaust the ... human mind." (Lucas, 'Minds, Machines and Gödel': 127)

Penrose would find no comfort in this anti-scientific conclusion—he wants there to be a fully scientific conception of the mind. But in his 1996 discussion 'Beyond the Doubting of a Shadow' Penrose says:

"[T]he Gödel diagonalization procedure can be applied to systems much more general than merely computational ones. Thus, my arguments would equally imply that our missing theory must be not just non-computational, but also beyond ... Turing's notion of oracle computation. ... It seems that the quality of 'understanding'—which is what

this discussion is effectively all about—is something very mysterious." (Penrose, 'Beyond the Doubting of a Shadow': 13.2)

Penrose's project was to find 'the missing science of consciousness', but in the end his ingenious and exhilarating discussion peters out with the platitude that the mind is 'very mysterious'.

What Did Gödel Think of the 'Gödel Argument'?

Gödel commenting on the philosophical significance of the incompleteness results:

"On ... the basis of what has been proved so far, it remains possible that there may exist (and even be empirically discoverable) a theorem-proving machine which in fact *is* equivalent to mathematical intuition, but cannot be *proved* to be so, nor even be proved to yield only *correct* theorems of finitary number theory." (Gödel in a note to Wang, 1972)

Gödel also says:

"The incompleteness results do not rule out the possibility that there is a theorem-proving computer which is in fact equivalent to mathematical intuition. ... If my result [incompleteness] is taken together with the rationalistic attitude which Hilbert had and which was not refuted by my results, then [we can infer] the sharp result that mind is not mechanical. This is so, because, if the mind were a machine, there would, contrary to this rationalistic attitude, exist number-theoretic questions undecidable for the human mind." (Gödel in conversation with Wang)

Gödel has hit the nail right on the head here. The incompleteness results by themselves certainly do not show that the mind is not a computer. The essential extra ingredient that must be added to the incompleteness results is the premiss of rationalistic optimism: the premiss that, as Hilbert famously put it, 'in mathematics there is no *ignorabimus*'—there are no mathematical questions that the human mind is incapable of settling, in principle at any rate, even if this is not so in practice.

Interestingly, Penrose appears to turn his back on the premiss of rationalistic optimism in the passage that we looked at earlier:

"[I]t need not be the case that human mathematical understanding is ... powerful enough, in principle, to solve each instance of the halting problem."

So Penrose ends up by jettisoning the mainspring of the Gödel argument!

Turing and Rationalistic Optimism

What did Turing have to say about rationalistic optimism?

Turing gave cautious expression to a form of rationalistic optimism when he said in 1936 (in the quotation discussed above):

"It is (so far as we know at present) possible that any assigned number of figures of δ can be calculated, but not by a uniform process." (Turing 1936)

He also said, in a lecture given circa 1951:

"By Gödel's famous theorem, or some similar argument, one can show that however the [theorem-proving] machine is constructed there are bound to be cases where the machine fails to give an answer, *but a mathematician would be able to.*" (Turing circa 1951, italics added)

And as we have already seen, in the quotation from the 1948 paper 'Intelligent Machinery' where Turing sets out the Mathematical Objection, he appears cautiously to endorse a version of rationalistic optimism, saying:

"On the other hand the human intelligence seems to be able to find methods of ever-increasing power for dealing with such problems[,] 'transcending' the methods available to machines." (Turing, 'Intelligent Machinery': 411)

In his 1950 paper 'Computing Machinery and Intelligence' Turing considers the obvious countermove against the Mathematical Objection, namely denying rationalistic optimism:

"The short answer to [the Mathematical Objection] is that although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect." (Turing, 'Computing Machinery and Intelligence': 451)

But rather than letting matters rest there, Turing continues:

"But I do not think [the Mathematical Objection] can be dismissed quite so lightly." (ibid.)

What Turing Might Have Said About the 'Sharp Result'?

Gödel: "If my result is taken together with the rationalistic attitude ... then [we can infer] *the sharp result that mind is not mechanical.* This is so, because, if the mind were a machine, there would, contrary to this rationalistic attitude, exist number-theoretic questions undecidable for the human mind."

What might Turing have said about what Gödel called the 'sharp result', had he commented on it? Even though Turing seems to have had some sympathy with rationalistic optimism, I think he would have wished to qualify Gödel's 'sharp result' considerably.

Gödel himself favoured immaterialism and (rather like Penrose and Lucas) tended towards mysticism about the mind. He said:

"Even if the finite brain cannot store an infinite amount of information, the spirit may be able to. The brain is a computing machine connected with a spirit." (Gödel in conversation with Wang)

This was certainly not Turing's way. Had he commented on Gödel's 'sharp result', Turing might have emphasised the point he made in his letter to Newman:

"If you think of various machines I don't see your difficulty. One imagines different machines allowing different sets of proofs ... " (Turing circa 1940)

Turing might have made the same retort to Penrose, too: *If you think of various machines I don't see your difficulty.*

Gödel's sharp result is just this: there is no *single* machine that is equivalent to mathematical intuition.

The Gödel argument is usually thought of as being a *reductio*. Assume that the mind is equivalent to a Turing machine, T say. Then the usual moves show that the mind cannot be equivalent to T; from which it is concluded that the mind is not a Turing machine. I'm suggesting that Turing's antidote to this is the idea that we think in terms of a dynamically changing mind, or collection of minds, each mind-stage being equivalent to a *different* Turing machine.

Turing uses the image of a sequence of increasingly powerful proof-producing machines in his 1950 paper 'Computing Machinery and Intelligence':

"In short, then, there might be men cleverer than any given machine, but then again there might be other machines cleverer again, and so on." (Turing, 'Computing Machinery and Intelligence': 451)

Turing's Apparent Answer to the Mathematical Objection

So it seems that Turing's answer to the Mathematical Objection goes something like this:

Pick any Turing machine T, then there may be a (developmental stage of) some mind M that is cleverer than T, but this has no tendency to show that (this stage of) M is not itself a Turing machine.—*It is perfectly consistent with the sharp result that this stage of M is a proof-producing Turing machine.*

So underlying this answer to the Mathematical Objection is what I call the multi-machine picture of mind:

The multi-machine picture of mind:

Human minds are Turing machines—in the sense that each developmental stage of a mind M is equivalent to some Turing machine, while different stages of M are equivalent to different Turing machines.

There is more to be said about the multi-machine picture, but before moving on to that let me mention the disagreement between myself and Andrew Hodges over whether Turing underwent a sea-change in his thinking about the mind.

The Pre- and Post-war Turing on the Mind

Andrew Hodges on Turing's 1939 paper 'Systems of Logic Based on Ordinals':

"the evidence is that at this time [Turing] was open to the idea that in moments of 'intuition' the mind appears to do something outside the scope of the Turing machine"

but:

"in the course of the war Turing dismissed the role for uncomputability in the description of mind, which once he had cautiously explored"

and:

"by 1945 Turing had come to believe computable operations had sufficient scope to include intelligent behaviour, and had firmly rejected the direction he had followed." (Hodges in his 1997 book *Turing: A Natural Philosopher*)

Hodges suggests that what changed Turing's mind was his experience at Bletchley Park:

"My guess is that there was a turning point in about 1941. After a bitter struggle to break U-boat Enigma, Turing could then taste triumph. Machines turned and people carried out mechanical methods unthinkingly, with amazing and unforeseen results. ... [I] suggest that it was at this period that [Turing] abandoned the idea that moments of intuition corresponded to uncomputable operations. Instead, he decided, the scope of the computable encompassed ... quite enough to include all that human brains did, however creative or original." (ibid.)

I have some comments from the mathematician Peter Hilton on these passages by Hodges. Hilton was Turing's colleague in Hut 8 at Bletchley Park and one of Turing's closest friends at this time.

"I must say that, if Alan Turing's thinking was undergoing so dramatic a change at that time, he concealed the fact very effectively." (Hilton, personal communication)

Hilton adds:

"I would never have said that we, working on Naval Enigma, 'carried out mechanical methods unthinkingly' (nor that our results were 'amazing and unforeseen')."

There is no textual evidence whatever for Hodges' postulated sea-change in Turing's thinking about the mind. What Turing said in his 1936 and 1939 papers is perfectly consistent with his post-war views. Nor do the letters to Newman from the Enigma period provide any evidence for a change of view—quite the reverse, in fact. Moreover, Turing's later work on the mind, far from

representing a rejection of his earlier ideas, appears to be a development of them.

Turing did not attempt to explain what he called the 'activity of the intuition', either in his 1939 paper nor the wartime letters to Newman. A human mathematician working according to the rules of a fixed logical system is in effect a proof-producing machine and when intuition supplies the mathematician with some new means of proof, he or she becomes a different proof-producing machine, capable of a larger set of proofs. How does the mathematician achieve this transformation from one proof-finding machine to another? The pre-war Turing was content to leave this question to one side, but the post-war Turing had a lot to say that is relevant to this question.

In his post-war writing on mind the term 'intuition' drops from view and what comes to the fore is the idea of *learning*—in the sense of devising or discovering—new methods of proof.

The Post-war Turing on Learning

In a lecture given circa 1951 Turing makes it clear that his—then radical—idea that machines can learn is the crux of his reply to the Mathematical Objection. He emphasises the importance of the idea of learning new methods of proof in his 1947 discussion of the Mathematical Objection too:

"[W]ith certain logical systems there can be no machine which will distinguish provable formulae of the system from unprovable ... Thus if a machine is made for this purpose it must in some cases fail to give an answer. On the other hand if a mathematician is confronted with such a problem he would search around and find new methods of proof, so that he ought eventually to be able to reach a decision about any given formula." (Turing in a lecture, 1947)

Turing's discussions of learning repeatedly emphasised:

- The importance of the learner making and correcting mistakes

"This danger of the mathematician making mistakes is an unavoidable corollary of his power of sometimes hitting upon an entirely new method." (Turing in a lecture, circa 1951)

- The importance of involving a random element

"[O]ne feature that ... should be incorporated ... is a 'random element'. ... This would result in the behaviour of the machine not being by any means completely determined by the experiences to which it was subjected." (Turing, *ibid.*)

- The importance of instruction modification (a matter underlined by Gualtiero in his discussion of the Mathematical Objection)

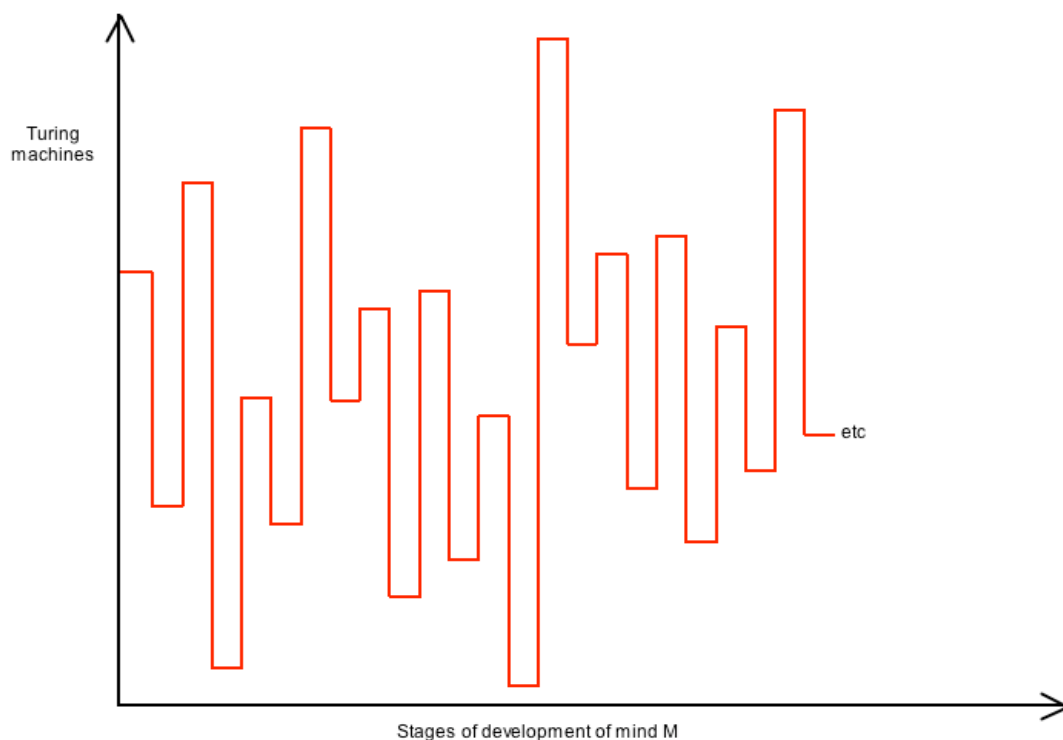
"What we want is a machine that can learn from experience. The possibility of letting the machine alter its own instructions provides the mechanism for this. ... One can imagine that after the machine had been operating for some time,

the instructions would have altered out of all recognition." (Turing in a lecture, 1947)

Thinking in terms of Turing machines, instruction-modification leads from one Turing machine to another: different instruction table, different Turing machine.

The Multi-Machine Picture of Mind

So here is a diagram of the multi-machine picture of mind. The learning mind successively mutates from one theorem-proving Turing machine into another. Idealising away death and other contingent resource-constraints, we can imagine the trajectory continuing indefinitely to the right:



Rationalistic Optimism and Uncomputability—Again

Is the function from mind-stages to Turing machines computable?—*Not if rationalistic optimism is true.*

So: if it is an open question whether some appropriate version of rationalistic optimism is true, then it is an open question whether this function is computable.

How could this function *fail* to be computable? Where could the uncomputability come from? Had Gödel commented specifically on the multi-machine picture of the human mind he might have said the following (he was actually commenting on the idea of a race of theorem-proving machines, analogous to the mind stages of the multi-machine picture):

"Such a state of affairs would show that there is something nonmechanical in the sense that the overall plan for the historical development of machines is not mechanical. If the general plan is mechanical, then the whole race can be summarised in one machine." (Gödel in conversation with Wang)

However, Gödel does not cash out this notion of an 'overall plan', and although his remark gets the issues into sharp focus, it does not really help very much with the specific question of where the uncomputability (if any) might come from.

We can only guess how Turing would have answered if questioned on this point. But given the emphasis that he placed on the inclusion of a random element in the learning process, he *might* have said:

The answer to your question is simple—the source of the uncomputability is randomness.

How is this learning of new methods of proof actually accomplished? Of course, Turing did not say. The question lies at the heart of the investigation of creativity.