*To appear in a special issue of* Kybernetes *on The Turing Test.*

# Plastic Machines: Behavioural Diversity and the Turing Test

**Michael Wheeler**
**Department of Philosophy**
**University of Stirling**

## Abstract

After proposing the Turing Test, Alan Turing himself considered a number of objections to the idea that a machine might eventually pass it. One of the objections discussed by Turing was that no machine will ever pass the Turing Test because no machine will ever "have as much diversity of behaviour as a man". He responded as follows: the "criticism that a machine cannot have much diversity of behaviour is just a way of saying that it cannot have much storage capacity". I shall argue that the objection cannot be dismissed so easily. The diversity exhibited by human behaviour is characterized by a kind of context-sensitive adaptive plasticity. Most of the time, human beings flexibly and fluently respond to what is relevant in a given situation. Moreover, ordinary human life involves an open-ended flow of shifting contexts to which our behaviour typically adapts in real time. For a machine to "have as much diversity of behaviour as a man" would be for that machine to keep its responses and behaviour relevant within such a flow. Merely giving a machine the capacity to store a huge amount of information and an enormous number of behaviour-generating rules will not achieve this goal. By drawing on arguments presented originally by Descartes, and by making contact with the frame problem in artificial intelligence, I shall argue that the distinctive context-sensitive adaptive plasticity of human behaviour explains why the Turing Test is such a stringent test for the presence of thought, and why it is much harder to pass than Turing himself may have realized.

Keywords: artificial intelligence; context; Descartes; frame problem

## 1. The Many Dimensions of the Turing Test

A human interrogator sits in front of a computer. She types in questions using the keyboard and reads responses (answers to her questions) from the screen. In the text on the screen, the two remote sources of these responses are labelled simply as X and Y. X and Y themselves are located out of direct sensory contact, in a different room. As it happens, one of them is a second human being, the other is a machine. But which is which? The interrogator's job is to decide this issue, and all she has to go on is the record of responses she receives to her questions. Her task is made non-trivial by the fact that it is the goal of the machine (or its designers) to fool her into making an incorrect identification. What I have just described is, of course, a stripped down version of the imitation game that constitutes the famous Turing Test, as introduced by Alan Turing in his classic paper, *Computing Machinery and Intelligence* (Turing 1950). It is a 'stripped down version' because it ignores Turing's own more complex set-up which involves: (a)

an initial configuration in which X and Y are a man and a woman, and in which the interrogator's job is to determine which is which, a task made non-trivial by the fact that it is the man's goal to fool the interrogator into making an incorrect identification; (b) a second stage in which the machine replaces the man, yielding the scenario already described; and (c) a criterion of success based on whether the number of correct identifications on the part of the interrogator is the same in both cases. As will become clear in a moment, depending on what we care about, these extra details may be important, but often they are not. What is beyond doubt, however, is that the Turing Test, in either form, constitutes an intellectually compelling proposal for how the products of the field known as artificial intelligence (AI) may be assessed. It is therefore unsurprising that it has gripped the imaginations of scientists and philosophers for over 50 years. But for what, exactly, is the Turing Test a test?

On one interpretation, the Turing Test is simply as a game to be played by smart AI-types, a game that presents an interesting, entertaining and potentially lucrative technical challenge, but one which, although it may inspire us to sing the praises of some creative computer programmers, has no broader consequences for how we think about thought and thinkers. This *deflationary* interpretation of the Turing Test will not be the focus of the present paper. Our concern will be with an *inflationary* interpretation, according to which the Turing Test embodies certain scientifically relevant, philosophically controversial, and morally charged views on what it takes for there to be thought and thinkers.

One reason for thinking in terms of 'deflationary' and 'inflationary' interpretations here is that Turing's own vision of the test might arguably be positioned somewhere between the two views just identified, although it strikes me that it is much closer to the latter. Consider: Turing starts his seminal paper by posing the question 'Can machines think?, which is undeniably a question with philosophical and moral aspects. He immediately argues, however, that this question is problematic, given that we don't have clear definitions of the terms 'machine' and 'think'. (Later in the paper he is more strident, describing his opening question as "too meaningless to deserve discussion" (Turing 1950, p.49).) He therefore proceeds to *replace* his original question with an alternative question, one that he describes as "closely related" to the original but "expressed in relatively unambiguous words" (Turing 1950, p.40). That new question is: will the interrogator decide wrongly as often when the imitation game is played between a computer and a woman as he or she does when the game is played between a man and a woman? Whatever revisions to our perspective on the Turing Test are mandated by the talk of 'replacement' here, the very fact that this new question is conceived by Turing as closely related to the one with which he starts out surely signals that his goal is not to divest his test of its philosophical importance, even though his aim is to side-step some tricky philosophical problems of definition.

One way to pursue an inflationary interpretation of the Turing Test is to hold that passing it constitutes a *sufficient* condition for the presence of thought. To bring this idea into view, consider the following question – call it the *sufficiency question*: if one found that there was no significant difference between the interrogator's success rate, whether the

scenario involves a lying man or a 'lying' machine, should one conclude that the machine genuinely thinks, in the same sense as the man? In other words, is an affirmative answer to Turing's replacement question enough to mandate an affirmative answer to his opening question? If so, then the differences between Turing's two questions are differences of scientific rigour, not philosophical force. So what should we say in answer to the sufficiency question? A reasonable answer to this question, as it is formulated above, might be, "well that depends". On what, you ask? On critical details such as the length of the test (the longer the fraud is successfully perpetrated, the better the evidence for genuine thought on the part of the machine), the profile of the interrogator (should it be someone who knows about psychology or AI, or someone who doesn't?), and so on. Let's say these sorts of details can be settled, such that we have a suitably specified version of the test. (For the rest of this paper, whenever I say 'Turing Test' I mean just such a suitably specified version.) Now let's ask the sufficiency question again: if one found that there was no significant difference between the interrogator's success rate, whether the scenario involved a lying man or a 'lying' machine, should one conclude that the machine thinks?

If you are inclined to answer 'yes' to the sufficiency question, you are probably being tempted by something like the following general principle, one that has application beyond the limits of the imitation game: where the observable outward behaviour of an artificial system is relevantly similar to the observable outward behaviour of a natural system, and where we already take the natural system to be a thinker, we have no good reason to withhold the status of genuine thinker from the artificial system. That's one way of pressing the point that exhibiting the right behaviour (in the case of Turing's imitation game, exhibiting the right linguistic behaviour) is sufficient for the presence of thought. On this view, anything that passes the Turing Test is, in a full and robust sense, a thinker – full stop, nothing more to be said. That's about as scientifically relevant (build me one), philosophically controversial (see below) and morally charged (we would have duties and responsibilities towards that thinker, and it would have duties and responsibilities towards us) as claims get.

It is probably fair to say that, among philosophers anyway, there is a good deal of scepticism directed at the claim that passing the Turing Test constitutes a sufficient condition for the presence of thought. This scepticism attracts what looks like a mixture of exasperation and impatience from some AI researchers. Thus, in a UK newspaper article (*The Observer*, 5th October 2008), the prominent cybernetician Kevin Warwick is quoted as saying: "I'm sure there will be philosophers who say, 'OK, [the machine has] passed the test, but it doesn't understand what it's doing'." We can add for ourselves what I take to be the unsaid semantic undercurrent: "it's passed the test – what else do these philosophers want?". So what fuels philosophical scepticism in this vicinity? In part, it is explained by a fear that the Turing Test (as we are interpreting it at the moment) invites a return to the bad old days of behaviourism in psychology and the philosophy of mind. But there are arguments too. Here is one of them.

In John Searle's famous Chinese Room thought experiment (Searle 1980), we take a mono-lingual English speaker, lock her in a room, and give her a rule book in English

that tells her how to manipulate a bunch of symbols that she can recognize only by their formal properties (e.g. their shapes). Then we pass a sequence of symbols into the room. Our room-trapped subject looks up that sequence in her rule book and then passes out another set of symbols, as determined by the appropriate rule. Unbeknownst to her, the symbols are Chinese characters. Chinese speakers outside the room interpret the first sequence of symbols as a particular question, and the second sequence as a sensible reply to that question. That's the hypothetical scenario. The question Searle asks is this: does our subject-in-the-room understand Chinese? The answer, it seems, is "no."

The most famous conclusion drawn by Searle on the basis of the Chinese room thought experiment concerns computation and thought. In her purely formal symbol-manipulating activity, the subject-in-the-room is the equivalent of a computer's central processing unit, and the rule book she follows is the equivalent of its program. So since merely performing formal symbol manipulations does not give our subject an understanding of Chinese, it cannot give a computer an understanding of Chinese. And this conclusion goes for mental states in general. So, the message goes, merely running the right program is not *sufficient* for thought (although it might, for all the thought experiment shows, be necessary). But now notice that, by hypothesis, the Chinese room system passes the Turing Test for Chinese linguistic behaviour. To native Chinese speakers, the behaviour of the Chinese room system is indistinguishable from that of a native Chinese speaker. Yet (our intuitions shout) the Chinese room system doesn't understand Chinese. This suggests a second conclusion, namely that *merely exhibiting the right behaviour is not sufficient for thought*. And that's in direct opposition to the idea that passing the Turing Test constitutes a sufficient condition for the presence of thought. As Searle himself puts it: "[P]recisely one of the points at issue [in the Chinese Room thought experiment] is the adequacy of the Turing-test. The example shows that there could be two "systems," both of which pass the Turing-test, but only one of which understands." (Searle 1980, p.74)

There are, then, established worries about the idea that, in passing the Turing Test, a machine would meet the sufficient conditions for the presence of thought. Even if these worries are justified, however, they say nothing about the idea that, in passing the Turing Test, a machine would meet a *necessary* condition for the presence of thought. And that's the dimension of the Turing Test on which I am going to focus for the rest of this paper, the idea that *being able to pass the Turing Test is a necessary condition for a machine to be a genuine thinker*. Taking this as my point of departure, I want to offer some reflections on why the Turing Test is in fact a *stringent* test for the presence of thought, and why it is in fact much harder to pass than Turing himself seems to have realized. In other words, I shall try to say why it has been so hard for AI systems to meet the necessary, let alone the sufficient, conditions for thought. In pursuing this issue, I want to stress that I am not in the business of defending any sort of in-principle claim that it is impossible for an artificial system to pass the Turing Test, and thus that Turing's original question 'Can machines think?' ought to be answered emphatically in the negative. No mere philosopher should be *that* bold. I think I can establish that the problem is hard, and say something about why it is hard. I shall even provide a philosophical perspective on some suggestive recent work in AI and neuroscience, in order to make some tentative

remarks on the general form that a solution to the problem might take. The rest I leave to history, or at least to what will be history, once AI has run its course.

## 2. The Behavioural Diversity Objection

In *Computing Machinery and Intelligence*, Turing himself considers a wide range of objections to the idea that a machine might eventually pass the Turing Test. A subset of these objections are collected together under the heading 'Arguments from Various Disabilities', and proceed by identifying a range of things that (according to Turing's imagined opposition) no mere machine will ever be able to do. These include various traits that human beings exhibit, either routinely or in particular cases, such as being kind, resourceful, beautiful or friendly, having initiative or a sense of humour, being able to tell right from wrong, making mistakes, falling in love, enjoying strawberries and cream, making some one fall in love with them, learning from experience, using words properly, being the subjects of their own thoughts, doing something really new, and – towards the end of Turing's list – the one that will concern us here, displaying behavioural diversity on a human scale. According to this objection, then, no machine will ever pass the Turing Test because no machine will ever "have as much diversity of behaviour as a man" (Turing 1950, p.53). Call this the *behavioural diversity objection*.

Turing's principal response to the behavioural diversity objection is swift. One gets the impression that he considers it to be decisive. He writes: "[the] criticism that a machine cannot have much diversity of behaviour is just a way of saying that it cannot have much storage capacity. Until fairly recently a storage capacity of even a thousand digits was very rare." (Turing 1950, p.55) To see how this response is supposed to work, consider the familiar image of a computational machine whose behavioural profile is determined by the ways in which it accesses, uses and manipulates internally represented bodies of information according to internally represented rules. Unpacked by way of this conceptualization of a machine, Turing's idea was that the impression that no such machine could have as much diversity of behaviour as a human being resulted from a kind of purblindness engendered in the casual observer by the limited amount of short-term and long-term information storage, and the limited number of rules, contained within the computational machines of his day. Of course, the storage capacity of present-day computers far outstrips those of Turing's own era, but that fact need not concern us here, because Turing's observation regarding the source of the purblindness in question is not what really makes his response tick. The key move (which is only implicitly suggested by Turing's actual text) comes next. If we give our computational machine an extremely large storage capacity, then it would be able to hold within it much more short-term and long-term information, and many many more behaviour-generating rules. By virtue of this massive increase in its access to information and rules, our machine would be endowed with the capacity to generate diverse behaviour on a par with the diversity of human behaviour, or at least there is no good reason to think that it couldn't. That's the crucial claim.

In spite of Turing's swift and confident response to the behavioural diversity objection, it seems to me that it cannot be dismissed so easily. Soon I shall explain why. Before that,

however, we need to consider, only to put aside, a second way in which Turing replies to the objection. Recall that I described Turing's appeal to increased storage capacity as his 'principal' response to the objection in question. I put things this way because he also remarks that the arguments from various disabilities are often disguised forms of a to-be-rejected line of reasoning that he calls "the argument from consciousness". In the present context, the key feature of the argument from consciousness is captured by the following comments from Turing: "Usually if one maintains that a machine can do one of these things [identified by the arguments from various disabilities], and describes the kind of method that the machine could use, one will not make much of an impression. It is thought that the method (whatever it may be, for it must be mechanical) is really rather base." (Turing 1950, p.56)

What is striking here is that the imagined purveyor of the argument from consciousness does not suggest that no mere machine could exhibit the behavioural phenomena in question (including behavioural diversity equivalent to that achieved by a human being). What she suggests is that even if we *could* specify a process by which a machine may exhibit those behavioural phenomena, that process will, by virtue of its mechanical character, be *the wrong sort* of process for us to conclude that the machine in question is actually thinking. In other words, exhibiting the behaviour in question is not *sufficient* for thought, because the crucial evidence for thought is provided not by behaviour but by having the right sort of inner process. Of course, the critic owes us a further argument here. We are considering something called "the argument from consciousness". So what's supposedly missing from any purely mechanistic process is conscious awareness. However, we haven't yet had an argument to the effect that such awareness cannot be realized by some sort of machine. Still, we can let that pass. What matters to us at the moment is that there is something distinctly odd about this second response by Turing to the behavioural diversity objection, given the way in which the arguments from various disabilities are originally presented. Here's why.

In their original form, the arguments from various disabilities identify behaviour that no machine could (allegedly) reproduce, not behaviour that, were a machine to reproduce it, would give us no conclusive evidence for the presence of thought. So seeing these arguments as disguised forms of the argument from consciousness divests them of their original character. In fact, it seems to me most natural to understand the arguments from various disabilities, in their original form, as identifying certain phenomena that are (allegedly) *necessary* for the presence of thought, and as claiming that no machine will ever reproduce those phenomena. In other words, the general claim on the table is that no mere machine will be able to meet the *necessary* conditions for the presence of thought. It is worth noting that the arguments from various disabilities, so conceived, *would* be disguised versions of a *different version* of the argument from consciousness to the one considered above. This would be a version which claimed that the behavioural phenomena in question (including behavioural diversity equivalent to that achieved by human beings) could not be reproduced by any non-conscious entity. If we consider some of the phenomena identified by the arguments from various disabilities as barriers to machine thought (e.g. falling in love, enjoying strawberries and cream), this interpretation is immediately plausible. It is rather more controversial, however, in the

target case of behavioural diversity, since it is not at all obvious that generating large-scale behavioural diversity is something that *only* a conscious entity could do. Since we are interested in behavioural diversity as a necessary but not a sufficient condition for the presence of thought, we can afford to ignore the thorny issue of consciousness and concentrate our attention on the explicit, undisguised version of the behavioural diversity objection. According to that version, the generation of behavioural diversity equivalent to that achieved by human beings is a necessary condition for the presence of thought, and no mere machine will be able to exhibit *that* degree of behavioural diversity. This is an idea with a history.

**3. Cartesian Machines**

In 1637, over 300 years before Turing described his imitation game, the great philosopher, mathematician and natural scientist Rene Descartes published one of his most important texts, namely the *Discourse on the Method of Rightly Conducting one's Reason and Seeking the Truth in the Sciences*, commonly known simply as the *Discourse* (Cottingham et al. 1985). Amazingly, within this seventeenth century text, Descartes reflects on the possibility of mechanizing thought. In the process he formulates and endorses a form of the behavioural diversity objection. Here is the key passage.

> [We] can certainly conceive of a machine so constructed that it utters words, and even utters words which correspond to bodily actions causing a change in its organs (e.g., if you touch it in one spot it asks you what you want of it, if you touch it in another it cries out that you are hurting it, and so on). But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to what is said in its presence, as the dullest of men can do... [And]... even though such machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they were acting not through understanding, but only from the disposition of their organs. For whereas reason is a universal instrument which can be used in all kinds of situations, these organs need some particular disposition for each particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act. (Cottingham et al. 1985, p.140)

So Descartes is clear that a machine might be built which is able to produce particular sequences of words as responses to specific stimuli, and, moreover, to perform individual task-specific actions as well as, if not better than, human beings. His negative claim is that no mere machine could either continually generate complex linguistic responses which are flexibly sensitive to varying contexts, in the way that all linguistically competent human beings do, or succeed in behaving appropriately in any context, in the way that all behaviourally normal human beings do. The way to understand this claim, I think (Wheeler 2005, 2008b), is to interpret the observations regarding language-use as identifying a particular case of a more general phenomenon. More specifically, the point that no machine (in virtue solely of its own intrinsic capacities) could reproduce the

generative and contextually sensitive linguistic capabilities displayed by human beings is actually just a local version of the more general point that no machine (in virtue solely of its intrinsic capacities) could reproduce the unrestricted range of adaptively flexible and contextually sensitive behaviour displayed by human beings. In other words, no mere machine could have as much diversity of behaviour as a man.

To see why Descartes' version of the behavioural diversity objection offers us an illuminating route into the issues, one needs to understand what is meant, in the target passage, by the term 'machine'. This will give us a definition of an entity that I shall call a *Cartesian machine*. (For detailed exegetical discussion that justifies this analysis, see Wheeler (2008b).) A Cartesian machine is a material system that (a) unfolds purely according to the laws of blind physical causation, (b) is susceptible to norms of correct and incorrect functioning, and (c) is either a special-purpose system or an integrated collection of special-purpose subsystems (where a system or subsystem is special-purpose if it is capable of producing appropriate actions only within some restricted task domain). Condition (a) identifies material systems whose behaviour can, for Descartes, be explained by the fundamental laws of mechanics. Condition (b) identifies the subset of such systems to which norms of correct and incorrect functioning are applicable. For example, a clock is a machine that has the function of telling the time. A broken clock fails to meet the norm of correct functioning for a machine of that kind, but of course it continues to follow the fundamental laws of mechanics just the same as if it were working properly. Condition (c) is designed to capture the feature of Cartesian machines that is to the fore in the target passage. We can bring this feature into proper view by considering a possible misinterpretation of Descartes' view. In the target passage, Descartes tells us that a machine acts "only from the disposition of [its] organs", organs that "need some particular disposition for each particular action". This choice of language may mislead us into thinking that, for Descartes, any entity which qualifies as a machine must be a look-up-table. However, we know from other things that Descartes says (again, see Wheeler 2008b for discussion) that he wants to make conceptual room for the idea that a machine may realize certain simple forms of locally driven intra-lifetime adaptation, learning and memory. The idea that a machine acts "only from the disposition of [its] organs" must therefore be interpreted in the light of this commitment. Of course, whatever forms of adaptation, learning and memory are present, they must take place within a system that needs "some particular disposition for each particular action". In my view, the best way to satisfy these dual constraints is to think of a Cartesian machine as a special-purpose system, or as an integrated collection of special-purpose subsystems (i.e. as meeting condition (c)). Look-up-tables are limiting cases of such systems, but the view allows that certain simple forms of locally driven intra-lifetime adaptation, learning and memory may be present.

Now that we understand what it is for something to be a Cartesian machine, we can see what is particularly interesting about Descartes' version of the behavioural diversity objection. Although he doesn't put the point in the way that I about to, the fact is that, according to Descartes, it is condition (c) that explains why the limits of machine intelligence lie where (he has argued) they do. If we concentrate on some individual, contextually-embedded human behaviour, it is possible that a Cartesian machine might be

built that incorporated a special-purpose mechanism (or set of special-purpose mechanisms) which would enable that machine to perform that behaviour as well as, or perhaps even better than, the human agent. However, the characteristic adaptive plasticity of human thought and behaviour is characterized by the fact that human beings keep their responses and behaviour relevant within an open-ended flow of shifting contexts. For a machine to "have as much diversity of behaviour as a man" would therefore be for that machine to realize a similar behavioral profile. But, Descartes claims, no mere machine could replicate human levels of behavioural plasticity. Why not? Because, he argues, it would be practically (i.e. empirically) impossible to incorporate into any one machine the vast number of special-purpose systems that would be required for that machine to consistently and reliably generate appropriate behaviour in all the different situations that make up an ordinary human life.

Now let's remind ourselves of Turing's principal response to the behavioural diversity objection, and ask ourselves how it stands in the light of Descartes' analysis. Turing argues that 'all' we need to do to meet the objection in question is to increase the storage capacity of the machine. Given Descartes' understanding of a machine as a special-purpose mechanism or as a set of special-purpose mechanisms, this amounts to the claim that one could build a machine with so many different special-purpose mechanisms that it had (more or less) one for every context into which it might be thrown. Descartes takes this to be practically impossible. His scientifically and philosophically informed empirical bet is that one simply couldn't stuff enough special-purpose mechanisms into one real physical machine. If Descartes is right, then no Cartesian machine – no set of special-purpose mechanisms – could display behavioural diversity on a human scale. This amounts to the claim that no Cartesian machine could pass the Turing Test. In other words, Descartes' claim is that no Cartesian machine could meet the necessary conditions for being a thinker.

**4. Neo-Cartesian Machines**

How do humans do it? According to Descartes, what machines lack, and what humans enjoy, is the faculty of understanding or reason, that "universal instrument which can be used in all kinds of situations". In other words, the distinctive and massive adaptive flexibility of human behaviour is explained by the fact that humans deploy non-mechanistic general-purpose reasoning processes. Now, between Descartes and contemporary AI came the birth of the digital computer. What this did (among other things) was to effect a widespread transformation in the very notion of a machine. To Descartes himself, reason, in all its (allegedly) general-purpose glory, looked staunchly resistant to mechanistic explanation. In the twentieth century, however, mainstream thinking in artificial intelligence was destined to be built (in part) on a concept that would no doubt have amazed and excited Descartes himself – the concept of a general-purpose reasoning machine, a *mechanical* "universal instrument which can be used in all kinds of situations". Mechanistic systems that realize general-purpose algorithms range from Newell and Simon's General Problem Solver (Newell and Simon 1963) to connectionist theories that think of the engine room of the mind as containing just a small number of general-purpose learning algorithms, such as Hebbian learning and back-propagation.

Such systems have shown us how general-purpose reason, that absolutely core and, according to Descartes, unmechanizable aspect of the Cartesian mind, might conceivably be realized by a machine. In the wake of such developments, we might now agree that human beings have the faculty of reason in Descartes' (general-purpose) sense – which is why they are capable of producing large-scale behavioural diversity – but hold that that crucial faculty can be mechanized. We might reasonably call such creations *neo-Cartesian machines*.

So do neo-Cartesian machines meet the behavioural diversity objection? I don't think so. For it is at this point that AI runs headlong into that long-standing irritant known as *the frame problem*. In its original form, the frame problem is the problem of how to characterize, using formal logic, those aspects of a state that are not changed by an action (see e.g. Shanahan 1997). However, the term has come to be used in a less narrow way, to name a multi-layered family of interconnected worries concerning the realization, retrieval and appropriate revision of epistemic or action-generating states (see e.g. the range of discussions in Pylyshyn 1987; see also Dennett 1984). The key questions are these: In a dynamically changing and open-ended world (rather than some artificially static and well-defined micro-world), how is a purely mechanistic system able to home in on just those aspects of all the things it senses, knows or believes are relevant in the present context of activity, while ignoring everything that is contextually irrelevant? How is that system then able to revise or act on that information in a contextually appropriate manner? In short, how might a 'mere' machine behave in ways that are adaptively sensitive to context-dependent relevance?

One first-pass response to these sorts of questions will be to claim that the machine should deploy stored heuristics (internally represented rules of thumb) that determine which of its rules and representations are relevant in the present situation. But are relevancy heuristics really a cure for the frame problem? It seems not. The processing mechanisms concerned would still face the problem of accessing just those relevancy heuristics that are relevant in the current context. So how does the system decide which of its stored heuristics are relevant? Another, higher-order set of heuristics would seem to be required. But then exactly the same problem seems to re-emerge at that processing level, demanding further heuristics, and so on. So, depending on how one looks at it, a combinatorial explosion or infinite regress beckons.

What conclusions should we draw from the existence and character of the frame problem? The frame problem provides us with another striking version of the behavioural diversity objection. To see why, note that it is at least arguable that the frame problem is a by-product of mind conceived as a neo-Cartesian machine, rather than of mind conceived as machine simpliciter. Consider: on the present proposal, what guarantees that "[mechanical] reason is [in principle] a universal instrument which can be used in all kinds of situations" is, at root, that the reasoning mechanism concerned has free and total access to a gigantic body of rules and information. Somewhere in that vast sea of structures lie the cognitive elements that are relevant to the present context. The perhaps insurmountable problem is how to find them in a timely fashion using a process of purely mechanical general-purpose search. If this is right, then we have evidence that no neo-

Cartesian machine – no general-purpose mechanism or set of general-purpose mechanisms – could display behavioural diversity on a human scale. This amounts to the claim that no neo-Cartesian machine could pass the Turing Test. In other words, no neo-Cartesian machine could meet the necessary conditions for being a thinker.

To my mind, we now have an explanation of why the Turing Test, conceived such that passing it would constitute meeting a necessary condition for the presence of thought, has proved to be such a stringent examination of the products of AI. Passing the Turing Test would mean building a machine that met the behavioural diversity objection. If we refuse Descartes' invitation to go beyond special-purpose mechanisms in our attempt to meet that objection, we are held in the jaws of Descartes' version of it. If, on the other hand, we accept Descartes' encouragement to go beyond special-purpose mechanisms, while maintaining that the resulting vision of general-purpose reason can be mechanized, we run headlong into the frame problem, and so confront an alternative and equally recalcitrant manifestation of the very same objection. Either way, the challenge is massive.

## 5. Cartesian Machines (Again)

In terms of our understanding of why the phenomenon of large-scale behavioural diversity makes the Turing Test so hard to pass, some progress is made by bringing the frame problem into view. As I have argued previously (Wheeler 2008a, forthcoming), the frame problem has two dimensions – intra-context and inter-context. In its *intra-context* dimension, the frame problem demands that we say how a purely mechanistic system might achieve appropriate, flexible and fluid action within a context. In its *inter-context dimension*, it demands that we say how a purely mechanistic system might achieve appropriate, flexible and fluid action in worlds in which adaptation to new contexts is open-ended and in which the number of potential contexts is indeterminate. One might think of the inter-context version of the frame problem as targeting the difficulty of explaining flexible, fluid and relevance-sensitive context-switching in an open-ended, dynamically changing environment. Now, if we think about the character of special-purpose mechanisms, that is, of Cartesian machines, we can see how the *intra-context* frame problem might be neutralized. One way of thinking about why it is that relevancy heuristics will (as I have suggested) fail to solve the frame problem is that they buy into an unpromising strategy of explicitly representing the context of activity in which the agent is, at any particular time, embedded. Hubert Dreyfus develops a similar point as follows:

> The significance to be given to each logical element [each internally represented piece of data] depends on other logical elements, so that in order to be recognized as forming patterns and ultimately forming objects and meaningful utterances each input must be related to other inputs by rules. But the elements are subject to several interpretations according to different rules and which rule to apply depends on the context. For a computer, however, the context itself can only be recognized according to a rule…

> …[T]o pick out two dots in a picture as eyes one must have already recognized the context as a face. To recognize this context as a face one must have distinguished its relevant features such as shape and hair from the shadows and highlights, and these, in turn, can be picked out as relevant only in a broader context, for example, a domestic situation in which the program can expect to find faces. This context too will have to be recognized by its relevant features, as social rather than, say, meteorological, so that the program selects as significant the people rather than the clouds. But if each context can be recognized only in terms of features selected as relevant and interpreted in terms of a broader context, the AI worker is faced with a regress of contexts. (Dreyfus 1992, pp.288-9)

One response to this sort of worry is that the explicit inner representation of context should be eschewed in favour of special-purpose mechanisms (or at least certain instances of special-purpose mechanisms) that implicitly define the context of activity in their basic operating principles. What does this mean? Here is a simple example that illustrates the point.

At any one time animals (including human beings), within any particular context of activity, do one thing rather than another, and what this is changes as the intra-context circumstances change. This is one version of what is known in the trade as the action selection problem. The traditional robotics approach to the action selection problem assumes the internal representation of appropriate behaviours and some sort of internal arbitration mechanism to decide between them. These features will in turn require a sensitivity to contextual relevance that will standardly be addressed using a rules and representations strategy. Rejecting this traditional strategy, Seth (1998) shows that, in a simple artificial world of power sources and traps, action selection desiderata such as prioritizing with respect to needs, sequencing behaviours appropriately, and opportunistic behaviour change, can be achieved by a minimal wheeled robot control architecture in which there are no internal representations of behaviour and no explicit arbitration procedures. Rather, a suite of independent artificially evolved activation functions directly link sensing and movement. The outputs (movement ‘recommendations’) from these special-purpose sensorimotor connections are numbers that are simply combined (roughly, summed and scaled) at the wheels as part of an ongoing perception-action cycle. In Seth’s solution to the action selection problem, context is not something that inner mechanisms must reconstruct in the form of inner representations, rules and heuristics, once those mechanisms have been triggered. Rather, context is something that is always there at the point of triggering, woven into the intrinsic fabric of the special-purpose mechanisms themselves. Thus, for these intrinsically context-dependent , special-purpose mechanisms, there is no intra-context frame problem. There is no intra-context frame problem because, once a mechanism is active, the kind of unmanageable search space that the frame problem places in the path of a purely general-purpose mechanism is simply never established. (To be clear, I am not claiming that only non-representational mechanisms may realize the key property of intrinsic context-dependence. In turning one’s back on representations of context, one need not turn one’s

back on representations altogether. For much more on this point, see Wheeler (2008a, forthcoming).)

So far so good. But even if there is reason to think that Cartesian machines may neutralize the intra-context dimension of the frame problem, the *inter-context* dimension of that problem remains. In other words, we have no answer to the question, 'What are the mechanistic principles by which a particular special-purpose mechanism (or suite of such mechanisms) is adaptively selected, in relevance-sensitive ways, from the vast range of such resources available to the machine?'. Thus imagine an agent entering a situation whose complexity places it beyond the reach of the sort of minimal action-selection solution deployed by Seth, and in which, due to the fact that what should be done is currently under-determined, more than one intrinsically context-dependent mechanism is poised to take charge of behaviour. If our only option at this point is to fall back on neo-Cartesian general-purpose reasoning mechanisms – mechanisms that, from a context-independent vantage point, survey the options and make the choice – we run straight back into the unwelcoming arms of frame problem. So have we exhausted the space of available options? Are there machines that are neither (wholly) Cartesian machines nor (wholly) neo-Cartesian machines (given that any particular machine may contain elements of one or both these other models)? If so, is there any evidence that such machines might defuse the inter-context frame problem? Having arrived at these questions, we are nearing the end of our investigation.

**6. Plastic Machines**

To pass the Turing Test, and thus (as I have presented things) to meet a necessary condition for the presence of thought, a machine would need to replicate the distinctive context-sensitive behavioural diversity that human beings display. This would mean taking the inter-context frame problem in its stride. Let's introduce the term *plastic machine* to label any machine that can do this. A plastic machine would meet the behavioural diversity objection. If human beings are machines, then we are plastic machines. So, in the wake of the arguments presented in the preceding sections, do we have any idea about what kinds of mechanistic processes might be at work in such machines? Let's bring things to a close with an all-too-brief attempt to answer this question.

Dreyfus (2008) is one thinker who has taken seriously the challenge presented to machine intelligence by (what I am calling) the inter-context frame problem. In describing how AI might make progress on this issue, Dreyfus draws lessons from Walter Freeman's neurodynamical models of brain activity (e.g. Freeman, 2000). Dreyfus writes:

> If Freeman is right… our sense of other potentially relevant familiar situations on the horizon of the current situation, might well be correlated with the fact that brain activity is not simply in one attractor basin at a time but is influenced by other attractor basins in the same landscape, as well as by other attractor landscapes which under what have previously been experienced as relevant conditions are ready to

> draw current brain activity into themselves. According to Freeman, what makes us open to the horizonal influence of other attractors is that the whole system of attractor landscapes collapses and is rebuilt with each new rabbit sniff [Freeman has worked extensively on rabbit olfaction], or in our case, presumably with each shift in our attention. And after each collapse, a new landscape may be formed on the basis of new significant stimuli – a landscape in which, thanks to past experiences, a different attractor is active. (Dreyfus 2008, p.360)

A machine that is "open to the horizonal influence of other attractors" in the way that Dreyfus describes would be a plastic machine. What Dreyfus offers us as a candidate for such a machine is a nonrepresentational dynamical system, one that is primed by past experience to pick up and enrich significance, a system whose constantly shifting attractor landscape is identified as causally explaining how newly encountered significances may interact with existing patterns of inner organization to create new global structures for interpreting and responding to stimuli. So what are we to make of this suggestion? There are reasons to think that what is on offer from Dreyfus-via-Freeman falls short of what is required (Wheeler forthcoming). The first thing to notice here is that it remains unclear from Dreyfus' text whether the crucial reconfiguration of the neural attractor landscape is supposed to be (i) *caused by* the attentional shift (as might be suggested by the parallel with the rabbit sniff and the talk of a new landscape being formed "on the basis of new significant stimuli") or (ii) *the causal basis of* the attentional shift (as might be suggested by the thought that the attractors in the landscape determine what we attend to). Either way, there is a worry. If (i) is the correct interpretation, then the shift in attention itself remains unexplained. But at least sometimes that shift in attention is presumably governed by, and thus presupposes, a grip on the way in which contexts of activity are changing. To that extent, then, Dreyfus' suggestion begs the question. On the other hand, if (ii) is the correct interpretation, then it seems that we are still owed an explanation of how it is that, out of all the attractors in the pre-transition landscape that have been significant in the past, and that might have become active, it is the relevant one that is ultimately selected. Once again, it seems, the key question is being begged.

In closing, it is worth repeating the second of these points, but with a different gloss. Elsewhere (Wheeler 2008a, forthcoming), I have suggested that Freeman's neurodynamical system realizes a form of causation that Andy Clark once dubbed *continuous reciprocal causation* (Clark 1997; for discussion, see e.g. Wheeler 2005). Continuous reciprocal causation is causation that involves multiple simultaneous interactions and complex dynamic feedback loops, such that (a) the causal contribution of each systemic component partially determines, and is partially determined by, the causal contributions of large numbers of other systemic components, and (b) those contributions may change radically over time. This species of causation, which is also to be found in some recent AI systems (e.g. the evolved GasNets deployed by Husbands et al. (1998); see Wheeler 2005 for philosophical discussion) plausibly bestows on a machine a certain kind of large-scale holistic flexibility, a flexibility that seems to be ripe to account, *in part*, for the fluid context-switching highlighted by the inter-context frame problem

(Wheeler 2005, 2008a, 2008b). Why do I say 'in part'? Because the fact that a machine may flexibly and holistically reconfigure itself on the basis of continuous reciprocal causation among its elements does not guarantee that the behaviours generated by that machine will remain contextually relevant. All that is assured is that the machine supports the kind of flexibility that, when harnessed appropriately (i.e. in context-sensitive ways), may help to generate fluid context-switching. In other words, although we may be in possession of *part of* the story about how plastic machines work, that story is still radically incomplete. To my mind, this is one core reason – perhaps *the* core reason – why the Turing Test, conceived such that to pass it would be to meet a necessary condition for the presence of thought, remains a formidable hurdle over which machine intelligence has not yet jumped. Put another way: as a response to the claim that a machine may one day pass the Turing Test, the behavioural diversity objection is a much better objection than Turing himself seems to have realized.

**Acknowledgments**

Some passages in this paper include textual material adapted from (Wheeler 2008b, forthcoming).

**References**

Clark, A. (1997), *Being There: Putting Brain, Body, And World Together Again*, MIT Press, Cambridge, Mass..

Cottingham, J., Stoothoff, R., & Murdoch, D. (Eds.). (1985), *The Philosophical Writings of Descartes, Vol. 1*, Cambridge University Press, Cambridge.

Dennett, D.C. (1984), "Cognitive wheels: the frame problem of AI, in Hookway, C. (Ed.), *Minds, Machines and Evolution: Philosophical Studies*, Cambridge University Press, Cambridge.

Dreyfus, H. L., (1992), *What Computers Still Can't Do: A Critique of Artificial Reason*, MIT Press, Cambridge, Mass..

Dreyfus, H. L. (2008), "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian", in Husbands, P., Holland, O. and Wheeler, M. (Eds.), *The Mechanical Mind in History*, MIT Press, Cambridge, Mass., pp.331-71. (A shortened version of this paper appears in under the same title in *Philosophical Psychology* Vol. 20, No. 2 (2007), pp.247-268. Another version appears under the same title in *Artificial Intelligence*, Vol. 171 (2007), pp.1137-1160. I have worked with the Husbands et al. (Eds.) version of the text, to which the cited page numbers refer.)

Husbands, P., Smith, T., Jakobi, N., and O'Shea, M. (1998), "Better living through chemistry: evolving GasNets for robot control", *Connection Science*, Vol. 10, pp.185-210.

Newell, A. & Simon, H.A. (1963), "GPS – a program that simulates human thought", in Feigenbaum, E.A.& Feldman, J. (Eds.), *Computers and Thought*. McGraw-Hill, New York, pp.279-96..

Pylyshyn, Z. (Ed.) (1987), *The Robot's Dilemma*, Ablex, Norwood, NJ.

Searle, J. R. (1980), "Minds, brains, and programs", *Behavioral and Brain Sciences*, Vol. 3, pp.417-24. Reprinted in Boden, M. A. (Ed.), (1990), *The Philosophy of Artificial Intelligence*, Oxford University Press, Oxford, pp. 67-88. Page numbers in the text refer to the reprinted version.

Seth, A. (1998), "Evolving action selection and selective attention without actions, attention or selection", in Pfeifer, P., Blumberg, B., Meyer. J.-A., and Wilson, S. (Eds.), *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, MIT Press, Cambridge, Mass., pp.139-46.

Shanahan, M. (1997), *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Intertia*, MIT Press, Cambridge, Mass.

Turing, A.M. (1950), "Computing machinery and intelligence", *Mind*, Vol. 59, pp.433-460. Reprinted in Boden M.A. (Ed.), (1990), *The Philosophy of Artificial Intelligence*, Oxford University Press, Oxford, pp. 40-66. Page numbers in the text refer to the reprinted version.

Wheeler, M. (2005), *Reconstructing The Cognitive World: The Next Step*, MIT Press, Cambridge, Mass.

Wheeler, M. (2008a), "Cognition in context: phenomenology, situated robotics and the frame problem", *International Journal of Philosophical Studies*, Vol. 16, No.3, pp.323-49.

Wheeler. M. (2008b), "God's machines: Descartes on the mechanization of mind", in Husbands, P., Holland, O. and Wheeler, M., (Eds.), *The Mechanical Mind in History*, MIT Press, Cambridge, Mass., pp.307-330.

Wheeler, M. (Forthcoming), "The problem of representation", in Gallagher, S. and Schmicking, D. (Eds.), *The Handbook of Phenomenology and Cognitive Science*, Springer, Berlin.