Draft of chapter that appears in Husbands, P., Holland, O. and Wheeler, M., eds. The Mechanical Mind in History, MIT Press, Cambridge, Mass.. 2008, pp.307-330.

God's Machines: Descartes on the Mechanization of Mind[1]

Michael Wheeler

1. Never Underestimate Descartes

In 1637 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. 1985a).[2] This event happened over 300 years before Turing, Ashby, Newell, Simon and the other giants of cybernetics and early artificial intelligence (AI) produced their seminal work. Approximately the same time-span separates the *Discourse* from the advent of the digital computer. Given these facts it will probably come as something of a surprise to at least some readers of this volume to discover that, in this text, Descartes reflects on the possibility of mechanizing mind. Not only that but, as I shall argue in this chapter, he elegantly identifies, and takes a far from anachronous or historically discredited stand on, a key question regarding the mechanization of mind, a question that, if we're honest with ourselves, we still don't really know how to answer. As I said, never underestimate Descartes.

2. Cartesian Machines

Before we turn to the key passage from the *Discourse* itself, we need to fill in some background. And to do that we need to understand what Descartes means by a 'machine'. In fact, given the different ways in which Descartes writes of machines and mechanisms, there are three things that he might mean by that term. They are:

a) a material system that unfolds purely according to the laws of blind physical causation;
b) a material system that is a machine in the sense of (a), but to which in addition certain norms of correct and incorrect functioning apply;
c) a material system that is a machine in the sense of (b), but which is also either (i) a special-purpose system or (ii) an integrated collection of special-purpose subsystems.[3]

As we shall see, Descartes thinks that while there are plenty of systems *in the actual world* that meet condition (a) alone, there is nothing *in the actual world* that meets condition (b) but not condition (c). Nevertheless, he thinks it is *conceivable* that something might meet (b) but not (c), so it is important to keep these two conditions distinct.

Let’s say that conditions (a), (b) and (c) define three different types of machine: type-a, type-b, and type-c respectively. So what sorts of things are there that count as type-a machines? Here the key observation (for our purposes) is that when it came to non-mental natural phenomena, Descartes was, for his time, a radical scientific reductionist. What made him so radical was his contention that (put crudely) biology was just a local branch of physics. Prior to Descartes, this was simply not a generally recognized option. The strategy had overwhelmingly been to account for biological phenomena by appealing to the presence of special vital forces, Aristotelian forms, or incorporeal powers of some kind. In stark contrast, Descartes argued that not only all the non-vital material aspects of nature, but also all the processes of organic bodily life – from reproduction, digestion, and growth, to what we would now identify as the biochemical and neurobiological processes going on in human and non-human animal brains – would succumb to explanations of the same fundamental character as those found in physics. But what was that character? According to Descartes, the distinctive feature of explanation in physical science was its wholly *mechanistic* nature. What matters here is not the details of one’s science of mechanics. In particular, nothing hangs on Descartes’ own understanding of the science of mechanics as being ultimately the study of nothing other than geometric changes in modes of extension.[4] What matters here is simply a general feature of mechanistic explanation, one shared by Descartes’ science of mechanics and our own, namely the view that in a mechanistic process, one event occurs after another, in a law-like way, through the relentless operation of blind physical causation. What all this tells us is that, for Descartes, the entire physical universe is ‘just’ one giant type-a machine. And that giant type-a machine consists of lots of smaller type-a machines, some of which are the organic bodies of non-human animals and human beings.

So far, so good. But when we say of a particular material system that it is a machine, we often mean something richer than that its behaviour can be explained by the fundamental laws of mechanics. We are judging additionally that certain norms of correct and incorrect functioning are applicable to that system. For example, a clock has the function of telling the time. A broken clock fails to meet that norm. Where such norms apply, the system in question is a type-b machine. To see how the introduction of type-b machines gives us explanatory leverage, we need note only that a broken type-b machine – a type-b machine that fails to function correctly judged against the relevant set of norms – continues to follow the fundamental laws of mechanics *just the same* as if it were working properly. A broken clock fails to perform its function of telling the time, but not by constituting an exception to the fundamental laws of mechanics. Thus we need the notion of a type-b machine, a machine as a *norm-governed* material system, to explain what changed about the clock, *as a machine*, when it stopped working. (Descartes himself makes these sorts of observations; see the *Sixth Meditation*; Cottingham et al. 1985b.)

It is a key feature of our understanding of the organic bodies of non-human animals and human beings – what I shall henceforth refer to as *bodily machines* or, to stress their generically shared principles of operation, as *the bodily machine* – that such systems count as machines in the richer, normatively loaded, type-b sense. This is essential to our understanding of health and disease. Thus a heart that doesn't work properly is judged to be failing to perform its function of pumping blood around the body. Descartes recognizes explicitly the normatively loaded character of the bodily machine. So where does he locate the source of the all-important norms of proper functioning? As Hatfield (1992) notes, Descartes vacillated over this point. Sometimes he seems to argue that all normative talk about bodily machines is, in truth, no more than a useful fiction in the mind of the observer, what he calls an "extraneous label". Thus, in the *Sixth Meditation*, he says: "When we say, then, with respect to the body suffering from dropsy, that it has a disordered nature because it has a dry throat but does not need a drink, the term 'nature' [the idea that the body is subject to norms of correct and incorrect functioning] is here used merely as an extraneous label" (Cottingham et al. 1985b, 69). At other times, however, an alternative wellspring of normativity presents itself. Descartes is clear that the bodily machine was designed by God. As he puts it in the *Discourse,* the body is a machine that was "made by the hands of God" (Cottingham et al. 1985a, 139). For Descartes, then, organic bodies, including those of human beings and non-human animals, are God's machines. Now, it seems correct to say that the functional normativity of a human-made machine is grounded in what the human designer of that artifact intended it to do. This suggests that the functional normativity of the bodily machine might reasonably be grounded in what its designer, namely God, intended it to do. Either way, the key point for our purposes is that some Cartesian machines, including all bodily machines, are explicable as norm-governed systems. Given the surely plausible thought that useful fictions can be explanatorily powerful, that would be true on either of Descartes' candidate views of the source of such normativity.[5]

Time to turn to the notion of a type-c machine – a machine as (additionally) a special-purpose system or as an integrated collection of special-purpose subsystems. To make the transition from type-b to type-c machines, we need to pay particular attention to the workings of the Cartesian bodily machine. A good place to start is with Descartes' account of the body's neurophysiological mechanisms.[6] According to Descartes, the nervous system is a network of tiny tubes along which flow the 'animal spirits', inner vapours whose origin is the heart. By acting in a way which (as Descartes himself explains it in the *Treatise on Man*) is rather like the bellows of a church-organ pushing air into the wind-chests, the heart and arteries push the animal spirits out through the pineal gland into pores located in various cavities of the brain (Cottingham et al. 1985a, 104). From these pores, the spirits flow down neural tubes that lead to the muscles, and thus inflate or contract those muscles to cause bodily movements. Of course, the animal spirits need to be suitably directed, so that the outcome is a bodily movement appropriate to the situation in which the agent finds herself. This is achieved in the following way. Thin nerve-fibres stretch from specific locations on the sensory periphery to specific locations in the brain. When sensory stimulation occurs in a particular organ, the connecting fibre tenses-up. This action opens a linked pore in the cavities of the brain, and thus releases a flow of animal spirits through a corresponding point on the pineal gland. Without further modification, this flow may be sufficient to cause an appropriate bodily movement. However, the precise pattern of the spirit-flow, and thus which behaviour actually gets performed, may depend also on certain guiding psychological interventions resulting from the effects of memory, the passions, and (crucially for what is to follow) reason.

The fine-grained details of Descartes' neurophysiological theory are, of course, wrong. However, if we shift to a more abstract structural level of description, what emerges from that theory is a high-level specification for a control architecture, one that might be realized just as easily by a system of electrical and biochemical transmissions – i.e., by a system of the sort recognized by contemporary neuroscience – as it is by Descartes' ingenious system of hydraulics. To reveal this specification let's assume that the bodily machine is left to its own devices (i.e. that it is left to function without the benefit of psychological interventions) and ask, 'What might be expected of it?'. As we have seen, Descartes describes the presence of dedicated links between (a) specific peripheral sites at which the sensory stimulation occurs, and (b) specific locations in the brain through which particular flows of movement-producing animal spirits are released. This makes it tempting to think that the structural organization of the unaided (by the mind) bodily machine would, in effect, be that of a look-up-table, a finite table of stored if-this-then-do-that transitions between particular inputs and particular outputs. This interpretation, however, ignores an important feature of Descartes' neurophysiological theory, one that we have not yet mentioned. The pattern of released spirits (and thus exactly which behaviour occurs) is sensitive to the physical structure of the brain. Crucially, as animal spirits flow through the neural tubes, they will sometimes modify the physical structure of the brain around those tubes, and thereby alter the precise effects of any future sensory stimulations. Thus Descartes clearly envisages the existence of locally acting bodily processes through which the unaided machine can, in principle, continually modify itself, so that its future responses to incoming stimuli are partially determined by its past interactions with its environment. The presence of such processes suggests that the bodily machine, on its own, is potentially capable of intra-lifetime adaptation, plus, it seems, certain simple forms of learning and memory. Therefore (on some occasions at least) the bodily machine is the home of mechanisms more complex than rigid look-up-tables.

What we need right now, then, is a high-level specification of the generic control architecture realized by the bodily machine, one that not only captures the intrinsic specificity of Descartes' dedicated mechanisms, but that also allows those mechanisms to feature internal states and intrinsic dynamics that are more complex than those of, for example, look-up-tables. Here is the suggestion: the bodily machine should be conceptualized as an integrated collection of special-purpose subsystems, where the qualifier 'special-purpose' indicates that each subsystem is capable of producing appropriate actions only within some restricted task-domain. Look-up-tables constitute limiting cases of such an architecture. More complex arrangements, involving the possibility of locally determined adaptive change within the task-domain, are, however, possible. What all this tells us is that, according to Descartes, the bodily machine is a type-c machine.

That concludes our brief tour of the space of Cartesian machines. Now, what about mechanizing the mind?

3. The Limits of the Machine

As we have seen, for Descartes, the phenomena of bodily life can be understood mechanistically. But did he think that the same mechanistic fate awaited the phenomena of mind? It might seem that the answer to this question must be a resounding ‘no’. One of the first things that anyone ever learns about Descartes is that he was a substance dualist. He conceptualized mind as a separate substance (metaphysically distinct from physical stuff) that causally interacts with the material world on an intermittent basis during perception and action. But if mind is immaterial, then (it seems) it can’t be a machine in any of the three ways that Descartes recognizes, since each of those makes materiality a necessary condition of machine-hood.

Game over? Not quite. Let’s approach the issue from a different angle, by asking an alternative question, namely ‘What sort of capacities might the bodily machine realize?’. Since the bodily machine is a type-c machine, this gives us a local (organism-centred) answer to the question ‘What sort of capacities might a type-c machine realize?’. One might think that the answer to this question must be autonomic responses and simple reflex actions (some of which may be modified adaptively over time), but not much else. If this is your inclination, then an answer that Descartes himself gives in the *Treatise on Man* might include the odd surprise, since he identifies not only “the digestion of food, the beating of the heart and arteries, the nourishment and growth of the limbs, respiration, waking and sleeping [and] the reception by the external sense organs of light, sounds, smells, tastes, heat and other such qualities”, but also “the imprinting of the idea of these qualities in the organ of the ‘common’ sense and the imagination, the retention or stamping of these ideas in the memory, the internal movements of the appetites and passions, and finally the external movements of all the limbs (movements which are... appropriate not only to the actions of objects presented to the sense, but also to the passions and impressions found in memory...)” (Cottingham et al. 1985a, 108). In the latter part of this quotation, then, Descartes takes a range of capacities that many theorists, even now, would be tempted to regard as psychological in character, and judges them to be explicable by appeal to nothing more fancy than the workings of the bodily machine. And here is another example of Descartes’ enormous faith in the power of ‘mere’ organic mechanism. According to Descartes, the first stage in the phenomenon of hunger is excitatory activity in certain nerves in the stomach. And he claims that this purely physical activity is sufficient to initiate bodily movements that are appropriate to food-finding and eating. Thus once again we learn from Descartes that the body, unaided by the mind, is already capable of realizing relatively complex adaptive abilities. (This is, of course, not the whole story about hunger. I’ll fill in the rest later.)

Should we be surprised by Descartes’ account of what the bodily machine can do? Not really. As we have seen, Descartes often appeals to artifacts as a way of illustrating the workings of the bodily machine. When he does this, he doesn’t focus on artifacts that, in his day, would have been thought of as dull or mundane, examples that might reasonably lead one to suspect that some sort of deflationary judgment on the body is in play. Rather, he appeals to examples that, in his day, would have been sources of popular awe and intellectual respect. These include clocks (rare, expensive and much admired as engineering achievements) and complex animal-like automata (as bought by the wealthy elite of seventeenth century Europe to entertain and impress even their most sophisticated guests). (For more on this, see Baker and Morris 1996, 92-3.) So when Descartes describes the organic body as a machine, we are supposed to gasp with admiration, not groan with disappointment. In fact we are supposed to be doubly impressed, since Descartes thought that the bodily machine was designed by God, and so is “incomparably better ordered than any machine that can be devised by man, and contains in itself movements more wonderful than those in any such machine” (*Discourse*, Cottingham et al. 1985a, 139). Our bodies are God’s machines and our expectations of them should be calibrated accordingly.

Now that we are properly tuned to Descartes’ enthusiasm for ‘mere’ mechanism, we can more reliably plot the limits that he placed on the bodily machine. Here, the standard interpretation of Descartes’ position provides an immediate answer: the bodily machine is incapable of conscious experience (see e.g. Williams 1990, 282-3). But is this really Descartes’ view? Departing from the traditional picture, Baker and Morris (1996) have argued that Descartes held some aspects of consciousness to be mechanizable. This sounds radical, until one discovers that, according to Baker and Morris, the sense in which, for Descartes, certain machines were conscious is the sense in which we can use expressions such as ‘see’ or ‘feel pain’ to designate “(the ‘input’ half of) fine-grained differential responses to stimuli (from both inside and outside the ‘machine’) mediated by the internal structure and workings of the machine” (p.99). Those who favour the traditional interpretation of Descartes might retaliate – with some justification I think, and in spite of protests by Baker and Morris (see pp.99-100) – that Descartes would not have considered this sort of differential responsiveness to stimuli to be a form of consciousness at all, at least not in any interesting or useful sense. Indeed, if he had thought of things in this way he would seemingly have been committed to the claim that all sorts of artifacts available in his day (e.g. the aforementioned entertainment automata) were conscious. It is very unlikely that he would have embraced such a consequence. Nevertheless, in spite of such worries about the Baker and Morris line, I think that some doubt has been cast on the thought that consciousness provides a sufficiently sharp criterion for determining where, on Descartes’ view, the limits of mere mechanism lie. It would be nice to find something better.

Time then to explore the passage from the *Discourse* in which Descartes explicitly considers the possibility of machine intelligence. Here it is:

> [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. 1985a, 140)

Once again Descartes' choice of language may mislead us into thinking that, on his view, any entity which qualifies (in the present context) as a machine must be a look-up-table. For example, he tells us that his imaginary robot acts "only from the disposition of [its] organs", organs that "need some particular disposition for each particular action". However, the way in which this robot is supposed to work is surely intended by Descartes to be closely analogous to the way in which the organic bodily machine is supposed to work. (Recall Descartes' enthusiasm for drawing illustrative parallels between the artificial and the biological when describing the workings of the bodily machine.) So we need to guarantee that there is conceptual room for Descartes' imaginary robot to feature the range of processes that, on his account, were found to be possible within the organic bodily machine. In other words, Descartes' imaginary robot needs to be conceived as an integrated collection of special-purpose subsystems, some of which may realize certain simple forms of locally driven intra-lifetime adaptation, learning and memory. In short, Descartes' robot is a type-c machine.

With that clarification in place, we can see the target passage as first plotting the limits of machine intelligence, and then explaining both why these limits exist and how human beings go beyond them. First let's see where the limits lie. Descartes argues that although a machine might be built which is (a) able to produce particular sequences of words as responses to specific stimuli, and (b) able to perform individual actions as well as, if not better than, human agents, no mere machine could either (c) continually generate complex linguistic responses which are flexibly sensitive to varying contexts, in the way that all linguistically competent human beings do, or (d) succeed in behaving appropriately in any context, in the way that all behaviourally normal human beings do. Here one might interpret Descartes as proposing two separate human phenomena – generative language-use and a massive degree of adaptive behavioural flexibility – both of which are beyond the capacities of any mere machine (for this sort of interpretation, see Williams 1990, 282-3). However, I think that there is another, perhaps more profitable way of understanding the conceptual relations in operation, according to which (a) and (c) ought to be construed as describing the special, linguistic instance of the general case described by (b) and (d). On this interpretation, although it is true that the human capacity for generative language-use is one way of marking the difference between mere machines and human beings, 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 restricted version of the 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. This alternative interpretation is plausible, I think, because when Descartes proceeds in the passage to explain why it is that no mere machine is capable of consistently reproducing human-level behaviour, he does not mention linguistic behaviour at all, but concentrates instead on the non-linguistic case.

To explain why the limits of machine intelligence lie where they do, Descartes argues as follows: Machines can act "only from the [special-purpose] disposition of their organs". Now, if we concentrate on some individual, contextually-embedded human behaviour, then it is possible that a machine might be built that incorporated a special-purpose mechanism (or set of special-purpose mechanisms) which would enable the machine to perform that behaviour as well as, or perhaps even better than, the human agent. However, it would be impossible to incorporate into any one machine the vast number of special-purpose mechanisms 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. So how do humans do it? 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 *general-purpose reasoning processes*.

It is important to highlight two features of Descartes' position here. First, Descartes' global picture is one in which, in human beings, reason and mechanism standardly work together to produce adaptive behaviour. To see this, let's return to the case of hunger introduced above. As I explained, the first stage in the phenomenon of hunger (as Descartes understands it) involves excitatory mechanical activity in the stomach that, in a way unaided by cognitive processes, initiates bodily movements appropriate to food-finding and eating. However, according to Descartes, some of the bodily changes concerned will often lead to mechanical changes in the brain which in turn cause associated ideas, including the conscious sensation of hunger, to arise in the mind. At this point in the flow of behavioural control, such ideas may prompt a phase of judgement and deliberation by the faculty of reason, following which the automatic movements generated by the original nervous activity may be revised or inhibited.

Second, the pivotal claim in Descartes' argument is that no single machine could incorporate the enormous number of special-purpose mechanisms that would be required for it to reproduce human-like behaviour. So what is the status of this claim? Descartes writes (in translation) that "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" (emphasis added). A lot turns on the expression 'for all practical purposes'. The French phrase in Descartes' original text is *moralement impossible* – literally 'morally impossible'. The idea that something which is morally impossible is something which is impossible *for all practical purposes* is defended explicitly by Cottingham (1992b, 249), who cites, as textual evidence, Descartes' explanation of moral certainty in the *Principles of Philosophy*. There the notion is unpacked as certainty that "measures up to the certainty we have on matters relating to the conduct of life which we never normally doubt, though we know it is possible absolutely speaking that they may be false" (Cottingham et al. 1985a, 290). I am persuaded by Cottingham's interpretation of the key phrase (despite the existence of alternative readings; see e.g. Baker and Morris, 1992, pp.183-8, especially footnote 331 on p.185). And I am equally persuaded by the use that Cottingham makes of that interpretation in his own discussion of the target passage from the *Discourse* (see Cottingham 1992b, pp.249-52). There he leans on his interpretation of *moralement impossible* to argue that Descartes' pivotal claim does not (according to Descartes anyway) have the status of a necessary truth. Rather, it is a scientifically informed empirical bet. Descartes believes that the massive adaptive flexibility of human behaviour cannot be generated or explained by the purely mechanistic systems of the body, since, as far as he can judge, it is practically impossible to construct a machine which contains enough different special-purpose mechanisms. However, he is, as far as this argument is concerned, committed to the view that the upper limits of what a mere machine might do must, in the end, be determined by rigorous scientific investigation and not by philosophical speculation. In other words, Descartes accepts that his view is a hostage to ongoing developments in science. And that explains why he thinks it *conceivable* (although, on the basis of present evidence, unlikely) that something might be a type-b machine without being a type-c machine.

4. Mechanics and Magic

Say one wanted to defend the view that mind may be mechanized *without exception*. How might one respond to Descartes' argument? Here is a potential line of argument. One might (a) agree that we have reason in Descartes' (general-purpose) sense, but (b) hold that reason (in that sense) can in fact be mechanized, and so (c) hold that the machines that explain human-level intelligence (general-purpose ones) are such as to escape Descartes' tripartite analysis of machine-hood. Let's see how one might develop this case.

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. According to Descartes' pre-computational outlook, machines simply were integrated collections of special-purpose mechanisms. To Descartes himself, then, 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, viz the concept of a *general-purpose reasoning machine*. The introduction of mechanistic systems that realize general-purpose reasoning algorithms is not something that Descartes himself even considered (how could he have?) but (one might argue) the arrival of such systems has shown how general-purpose reason, that absolutely core and, according to Descartes, unmechanizable aspect of the Cartesian mind, might conceivably be realized by a bodily machine. Let's call such a machine a type-d machine. Evidence of the importance of type-d machines to AI abounds in the literature. It includes massively influential individual models, such as Newell and Simon's *General Problem Solver* (Newell and Simon 1963), a program that used means-end reasoning to construct a plan for systematically reducing the difference between some goal-state (as represented in the machine) and the current state of the world (as represented in the machine). And it includes generic approaches to machine intelligence, such as mainstream connectionist theories (more on which below) 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.[7]

So is this a good response to Descartes' argument? I don't think so. Why? Because it runs headlong into a long-standing enemy of AI known as the *frame problem*. In its original form, the frame problem is the problem of characterizing, 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 to do with the updating of epistemic states in relevance-sensitive ways (see e.g. the range of discussions in Pylyshyn 1987). A suitably broad definition is proposed by Fodor, who describes the frame problem as "the problem of putting a "frame" around the set of beliefs that may need to be revised in the light of specified newly available information" (Fodor 1983, 112-13). Here I shall be concerned with the frame problem in its more general form.

To see why the framing requirement described by Fodor constitutes a bona fide problem, as opposed to merely a description of what needs doing, consider the following example due to Dennett (1984). Imagine a mobile robot that has the capacity to reason about its world by proving theorems on the basis of internally stored, logic-based representations. (This architecture is just one possibility. Nothing about the general frame problem means that it is restricted to control systems whose representational states and reasoning algorithms are logical in character.) This robot needs power to survive. When it is time to find a power-source, the robot proves a theorem such as PLUG-INTO(Plug, Power-source). The intermediate steps in the proof represent sub-goals which the robot needs to achieve, in order to succeed at its main goal of retrieving a power-source (cf. the means-end reasoning algorithm deployed by GPS, as mentioned above).

Now, consider what might happen when our hypothetical robot is given the task of collecting its power-source from a room which also contains a bomb. The robot knows that the power-source is resting on a wagon, so it decides (quite reasonably, it seems) to drag that wagon out of the room. Unfortunately the bomb is on the wagon too. The result is a carnage of nuts, bolts, wires, and circuit boards. It is easy to see that the robot was unsuccessful here because it failed to take account of one crucial side-effect of its action, viz the movement of the bomb. So, enter a new improved robot. This one operates by checking for every side-effect of every plan that it constructs. This robot is unsuccessful too, simply because it never gets to perform an action. It just sits there and ruminates. What this shows is that it is no good checking for every side-effect of every possible action before taking the plunge and doing something. There are just too many side-effects to consider, and most of them will be entirely irrelevant to the context of action. For example, taking the power-source out of the room changes the number of objects in the room, but, in this context, who cares? And notice that the robot needs to consider not only things about its environment which have changed, but also things which have not. Some of these will be important some of the time, given a particular context. So the robot needs to know which side-effects of its actions and which unchanged facts about its world are relevant, and which are not. Then it can just ignore all the irrelevant facts. Of course, if the context of action changes, then what counts as relevant may change. For instance, in a different context, it may be absolutely crucial that the robot takes account of the fact that, as a result of its own actions, the number of objects in the room has changed.

We have just arrived at the epicentre of the frame problem, and it's a place where the idea of mind as machine confronts a number of difficult questions. For example, given a dynamically changing world, how is a purely mechanistic system to take account of those state-changes in that world (self-induced or otherwise) which matter, and those unchanged states in that world which matter, whilst ignoring those which do not? And how is that system to retrieve and (if necessary) to revise, out of all the beliefs that it possesses, just those beliefs that are relevant in some particular context of action? In short, how might a 'mere' machine behave in ways that are 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 (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. It is not merely that some sort of combinatorial explosion or infinite regress beckons here (which it does). A further concern, in the judgment of some notable authorities, is that we seem to have no good idea of how a computational process of relevance-based update might work. As Horgan and Tienson (1994) point out, the situation cannot be that the system first retrieves an inner structure (an item of information or a heuristic), and then decides whether or not it is relevant, as that would take us back to square one. But then how can the system assign relevance until the structure has been retrieved?

But if the frame problem is *such* a nightmare, how come AI hasn't simply ground to a halt? According to many front-line critics of the field (including Dreyfus, this volume), most AI researchers (classical and connectionist) have managed to side-step the frame problem precisely because they have tended to assume that real-world cognitive problem-solving can be treated as a kind of messy and complicated approximation to reasoning (or learning) in artificially restricted worlds that are relatively static, essentially closed, and feature some small number of contexts of action. In such worlds, all the contexts that could possibly arise may be identified and defined, alongside all the factors that could possibly count as relevant within each of them. So the programmer can either take comprehensive and explicit account of the effects of every action or change, or work on the assumption that nothing changes in a scenario unless it is explicitly said to change by some rule. And if those strategies carried too high an adaptive cost in terms of processing resources, well-targeted relevancy heuristics would appear to have a good chance of heading off the combinatorial explosions and search difficulties that threaten.

One might think, however, that the actual world often consists of an indeterminate number of dynamic, open-ended, complex scenarios in which context-driven and context-determining change is common and ongoing, and in which vast ranges of cognitive space might, at any time, contain the relevant psychological elements. It is in this world that the frame problem really bites, and in which (it seems) the aforementioned strategies must soon run out of steam.

From what we have seen so far, the frame problem looks to be a serious barrier to the mechanization of mind. Indeed, one possible conclusion that one might draw from the existence and nature of the frame problem is that human intelligence is a matter of magic not mechanics. However, it is at least arguable that the frame problem is in fact a by-product of mind conceived as a general-purpose (type-d) machine, rather than of mind conceived as machine *simpliciter*. What mandates this less extreme conclusion? It's the following line of thought. 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 search. What this suggests is that we might do well to reject the very idea of the bodily machine as a general-purpose reasoning machine, and to investigate what happens to the frame problem if we refuse to accept Descartes' invitation to go beyond special-purpose mechanisms in our understanding of intelligence.

Here is the view from the armchair: a system constructed from a large number of special-purpose mechanisms will simply take the frame problem in its stride. This is because, in any context of action, the special-purpose mechanism that is appropriately activated will, as a direct consequence of its design, have access to no more than a highly restricted subset of the system's stock of rules and representations. Moreover, that subset will include just those rules and representations that are relevant to the adaptive scenario in which the system finds itself. Therefore the kind of unmanageable search space that the frame problem places in the path of a general-purpose mechanism is simply never established. So those are the armchair intuitions. But is there any evidence to back them up? Here is a much-discussed model from the discipline of biorobotics.

Consider the ability of the female cricket to find a mate by tracking a species-specific auditory advertisement produced by the male. According to Barbara Webb's robotic model of the female cricket's behaviour, here, roughly, is how the phonotaxis system works (for more details, see Webb 1993, 1994, or the discussion in Wheeler 2005). The basic anatomical structure of the female cricket's peripheral auditory system is such that the amplitude of her ear-drum vibration will be higher on the side closer to a sound-source. Thus, if some received auditory signal is indeed from a conspecific male, all the female needs to do to reach him (all things being equal) is to continue to move in the direction indicated by the ear-drum with the higher amplitude response. So how is that the female tracks only the correct stimulus? The answer lies in the activation profiles of two interneurons (one connected to each of the female cricket's ears) that mediate between ear-drum response and motor behaviour. The decay rates of these interneurons are tightly coupled with the specific temporal pattern of the male's song, such that signals with the wrong temporal pattern will simply fail to produce the right motor-effects.

Why is this robotic cricket relevant to the frame problem? The key idea is suggested by Webb's own explanation of why the proposed mechanism is adaptively powerful: "Like many other insects, the cricket has a simple and distinctive cue to find a mate, and consequently can have a sensory-motor mechanism that works for this cue and nothing else: there is no need to process sounds in general, provided this specific sound has the right motor effects. Indeed, it may be advantageous to have such specificity built in, because it implicitly provides 'recognition' of the correct signal through the failure of the system with any other signal" (Webb 1993, 1092). A reasonable gloss on this picture would be that, rather than starting outside of context and having to find its way in using relevancy heuristics and so on, the cricket's special purpose mechanism, in the very process of being activated by a specific environmental trigger, brings a context of activity along with it, implicitly realized in the very operating principles which define that mechanism's successful functioning. Thus, to repeat the armchair intuition, there is no frame problem here because the kind of unmanageable search space that the frame problem places in the path of a general-purpose mechanism is simply never established.

If one takes the sort of mechanism described by Webb, generalizes the picture so that one has an integrated architecture of such mechanisms, and then looks at the result through historically tinted glasses, then it seems to reflect two of Descartes' key thoughts: (a) that organic bodies are collections of special-purpose subsystems (type-c machines), and (b) that such subsystems (individually and in combination) are capable of some pretty fancy adaptive stuff. Moreover, this would seem to be a machine that solves the frame problem (in effect, by not letting it arise). This looks to be a step forward – and it is. Unfortunately, however, it falls short of what we need. It falls short because while it solves the frame problem, it doesn't solve Descartes' problem. As we know, Descartes himself argued that there was a limit to what any collection of special-purpose mechanisms could do: no single machine, he thought, could incorporate the enormous number of special-purpose mechanisms that would be required for it to reproduce the massive adaptive flexibility of human behaviour. That's why, in the end, Descartes concludes that intelligent human behaviour is typically the product of general-purpose reason. Nothing we have discovered so far suggests that Descartes was wrong about that. Here's the dilemma, in a nutshell: If we mechanize general-purpose reason, we get the frame problem; so that's no good. But if we don't mechanize general-purpose reason, we have no candidate mechanistic explanation for the massive adaptive flexibility of human behaviour; so that's no good either. The upshot is that if we are to resist Descartes' anti-mechanistic conclusion, something has to give.

At this juncture let's return to the target passage from the *Discourse*. There is, I think, a tension hidden away in Descartes' claim that (as it appears in the standard English translation) "reason is a universal instrument which can be used in all kinds of situations". Strictly speaking, if reason is a *universal* instrument then, at least potentially, it ought to be possible for it to be applied unrestrictedly, across the cognitive board. If this is right then ‘all kinds of situations’ needs to be read as ‘any kind of situation’. However, I don’t think we ordinarily use the phrase ‘all kinds of’ in that way. When we say, for example, that the English cricket team, repeatedly slaughtered by Australia during the 2006-7 Ashes tour, are currently in ‘all kinds of problems’, we mean not that the team faces all the problems there are in the world, but rather that they face a wide range of different problems. But now if this piece of ordinary language philosophy is a reliable guide for how we are meant to read Descartes’ claim about reason, then that claim is weakened significantly. The suggestion now is only that reason is an instrument that can be used in *a wide range of different situations*.

With this alternative interpretation on the table, one might think that the prospects for an explanation of human reason in terms of the whirrings of a type-c machine are improved significantly. The argument would go like this:

> Human reason is, in truth, a suite of specialized psychological skills and tricks with domain-specific gaps and shortcomings. That would still be an instrument that can be used in a wide range of different situations. And by Descartes’ own lights, a material system of integrated special-purpose mechanisms (a type-c Cartesian machine) ought to be capable of this sort of cognitive profile.

But this is to move too quickly. For even if the claim that reason is a “universal instrument” over-states just how massively flexible human behaviour really is, it’s undeniably true that human beings are impressively flexible. Indeed, the provisional argument just aired fails to be sufficiently sensitive to the thought that an instrument that really can be used successfully across a wide range of different situations is an instrument that must be capable of fast, fluid and flexible context-switching. Crucially, this sort of capacity for real-time adaptation to new contexts appears to remain staunchly resistant to exhaustive explanation in terms of any collection of *purely* special-purpose mechanisms. The worry is this: so far, we have no account of the mechanistic principles by which a particular special-purpose mechanism is selected from the vast range of such mechanisms available to the agent and then placed in control of the agent’s behaviour at a specific time. One can almost hear Descartes’ ghost as he claims that we will ultimately need to posit a general-purpose reasoning system whose job it is to survey the options and make the choice. But if that’s the ‘solution’, then the door to the frame problem would be re-opened, and we would be back to square one (or thereabouts).

5. Plastic Machines

Our task, then, is to secure adaptive flexibility on a scale sufficient to explain open-ended adaptation to new contexts, without going beyond mere mechanism and without a return to Cartesian general purpose reason. Here is a suggestion (an incomplete one, I freely admit) for how this might be achieved.

Roughly speaking, the term ‘connectionism’ picks out research on a class of intelligent machines in which a (typically) large number of interconnected units process information in parallel. In as much as the brain too is made up of a large number of interconnected units (neurons) that process information in parallel, connectionist networks are ‘neurally inspired’, although usually at a massive level of abstraction. Each unit in a connectionist network has an activation level regulated by the activation levels of the other units to which it is connected, and, standardly, the effect of one unit on another is either positive (if the connection is excitatory) or negative (if the connection is inhibitory). The strengths of these connections are known as the network’s weights, and it is common to think of the network’s ‘knowledge’ as being stored in its set of weights. The values of these weights are (in most networks) modifiable, so, given some initial configuration, changes to the weights can be made which improve the performance of the network over time. In other words, within all sorts of limits imposed by the way the input is encoded, the specific structure of the network, and the weight-adjustment algorithm, the network may learn to carry out some desired input-output mapping.

Most work on connectionist networks has tended to concentrate on architectures that, in effect, limit the range and complexity of possible network dynamics. These features include (a) neat symmetrical connectivity, (b) noise-free processing, (c) update properties which are based either on a global, digital pseudo-clock or on methods of stochastic change, (d) units which are uniform in structure and function, (e) activation passes that proceed in an orderly feed-forward fashion, and (f) a model of neurotransmission in which the effect of one neuron’s activity on that of a connected neuron will simply be either excitatory or inhibitory, and will be mediated by a simple point-to-point signalling process. Quite recently, however, some researchers have come to favour a class of connectionist machines with richer system dynamics, so-called *dynamical neural networks* (henceforth *DNNs*).

What we might, for convenience, call mark-one DNNs feature the following sorts of properties (although not every bona fide example of a mark-one DNN exhibits all the properties listed): asynchronous continuous-time processing, real-valued time delays on connections, non-uniform activation functions, deliberately introduced noise, and connectivity which is not only both directionally unrestricted and highly recurrent, but also not subject to symmetry constraints (see e.g. Beer and Gallagher 1992, Husbands et al. 1995). Mark-two DNNs add two further twists to the architectural story. In these networks, christened GasNets (Husbands et al. 1998), the standard DNN model is augmented with (i) modulatory neurotransmission (according to which fundamental properties of neurons, such as their activation profiles, are transformed by arriving neurotransmitters), and (ii) models of neurotransmitters that diffuse virtually from their source in a cloud-like, rather than a point-to-point, manner, and thus affect entire volumes of processing structures. GasNets thus provide a platform for potentially rich interactions between two interacting and intertwined dynamical mechanisms – virtual cousins of the electrical and chemical processes in real nervous systems. Diffusing ‘clouds of chemicals’ may change the intrinsic properties of the artificial neurons, thereby changing the patterns of ‘electrical’ activity, whilst ‘electrical’ activity may itself trigger ‘chemical’ activity. So, dropping the scare quotes, these biologically inspired machines feature neurotransmitters that may not only transform the transfer functions of the neurons on which they act, but which may do so on a grand scale, as a result of the fact that they act by gaseous diffusion through volumes of brain-space, rather than by electrical transmission along connecting neural wires.

Systems of this kind have been artificially evolved[8] to control mobile robots for simple homing and discrimination tasks. So what does the analysis of such machines tell us? Viewed as static wiring diagrams, many of the successful GasNet controllers appear to be rather simple structures. Typical networks feature a very small number of primitive visual receptors, connected to a tiny number of inner and motor neurons by just a few synaptic links. However, this apparent structural simplicity hides the fact that the dynamics of the networks are often highly complex, involving, as predicted, subtle couplings between chemical and electrical processes. For example, it is common to find adaptive use being made of oscillatory dynamical sub-networks, some of whose properties (e.g., their periods) depend on spatial features of the modulation and diffusion processes, processes which are themselves determined by the changing levels of electrical activity in the neurons within the network (for more details, see Husbands et al. 1998). Preliminary analysis suggests that these complex interwoven dynamics will sometimes produce solutions which are resistant to any modular decomposition. However, there is also evidence of a kind of transient modularity in which, over time, the effects of the gaseous diffusible modulators drive the network through different phases of modular and non-modular organization (Husbands, personal communication).

What seems clear, then, is that the sorts of machines just described realize a potentially powerful kind of ongoing fluidity, one that involves the functional and even the structural reconfiguration of large networks of components. This is achieved on the basis of bottom-up systemic 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, moreover, (b) those contributions may change radically over time. (This is what Clark (1997) dubs continuous reciprocal causation.) At root, GasNets are mechanisms of significant adaptive plasticity, and it seems plausible that it is precisely this sort of plasticity that, when harnessed and tuned appropriately by selection or learning to operate over different time-scales, may be the mechanistic basis of open-ended adaptation to new contexts. It is a moot point whether or not this plasticity moves us entirely beyond the category of type-c machines. To the extent that one concentrates on the way in which GasNets may shift from one kind of modular organization to another (in realizing the kind of transient modularity mentioned above), the view is compatible with a story in which context switching involves a transition from one arrangement of special-purpose systems to another. Under these circumstances, perhaps it would be appropriate to think of GasNets as type-c.5 machines.

6. Concluding Remarks

In the *Discourse,* Descartes lays down a challenge to the advocate of the mechanization of mind. How can the massive adaptive flexibility of human-level intelligence be explained without an appeal to a non-mechanistic faculty of general-purpose reason?

Descartes' scientifically informed empirical bet is that it cannot. Of course, his conclusion is based on an understanding of machine-hood that is linked conceptually to the notion of special-purpose mechanisms. This understanding, and thus his conclusion, has been disputed by the subsequent attempt in AI to mechanize general-purpose reason. However, since this ongoing attempt is ravaged by the frame problem, it does not constitute a satisfactory response to Descartes' challenge. Are plastic machines, as exemplified by GasNets, the answer? As I write, I know of no empirical work which demonstrates conclusively that the modulatory processes instantiated in GasNets can perform the crucial context-switching function that I have attributed to them. For while there is abundant evidence that such processes can mediate the transition between different phases of behaviour within the same task (Smith et al. 2001), that is not the same thing as switching between contexts, which typically involves a re-evaluation of what the current task might be. Nevertheless, it is surely a thought worth pursuing that fluid functional and structural reconfiguration, driven in a bottom-up way by low-level neuro-chemical mechanisms, may be at the heart of the more complex capacity. But that is my scientifically informed empirical bet, and it is one that needs to be balanced against Descartes' own. At present Descartes' challenge remains essentially unanswered. Never underestimate Descartes. (Have I said that?)

References

Baker, G. & Morris, K.J. (1996). *Descartes' Dualism*. Routledge, London and New York.

Beer, R.D. & Gallagher, J.G. (1992). Evolving dynamic neural networks for adaptive behavior. *Adaptive Behavior*, 1, 91-122.

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

Cottingham, J. (Ed.). (1992a). *The Cambridge Companion to Descartes*. Cambridge University Press, Cambridge.

Cottingham, J. (1992b). Cartesian dualism: Theology, metaphysics, and science. In (Cottingham 1992a), 236-57.

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

Cottingham, J., Stoothoff, R., & Murdoch, D. (Eds.). (1985b). *The Philosophical Writings of* Descartes, Vol. 2. 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.

Dennett, D.C. (1987). *The Intentional Stance*. MIT Press/ Bradford Books, Cambridge, Mass..

Dreyfus, H.L. This volume. Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian.

Fodor, J.A. (1983). *The Modularity of Mind*. MIT Press / Bradford Books, Cambridge, Mass.

Hatfield, G. (1992). Descartes' physiology and its relation to his psychology. In Cottingham (1992a), 335-70.

Horgan, T. & Tienson, J. (1994). A nonclassical framework for cognitive science. *Synthese*, 101, 305-45.

Husbands, P., Harvey, I., & Cliff, D. (1995). Circle in the round: State space attractors for evolved sighted robots. *Robotics and Autonomous Systems*, 15, 83-106.

Husbands, P., Smith, T., Jakobi, N., & O'Shea, M. (1998). Better living through chemistry: Evolving GasNets for robot control. *Connection Science*, 10(3/4), 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, 279-96..

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

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

Smith, T., Husbands, P., & O'Shea, M. (2001). Neural networks and evolvability with complex genotype-phenotype mapping. In Kelemen, J. and Sosik, P. (Eds.), 2001, *Advances in Artificial Life: Proceedings of the Sixth European Conference on Artificial Life,* Berlin and Heidelberg. Springer-Verlag, 272-81.

Webb, B. (1993). Modeling biological behaviour or 'dumb animals and stupid Robots'. In *Pre-Proceedings of the Second European Conference on Artificial Life,* 1090-103.

Webb, B. (1994). Robotic experiments in cricket phonotaxis. In Cliff, D., Husbands, P., Meyer, J.-A., & Wilson, S.W. (Eds.). 1994. *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, Cambridge, Mass. MIT Press / Bradford Books, 45-54.

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

Williams, B. (1990). *Descartes: The Project of Pure Enquiry*. Penguin, London.

[1] This chapter draws extensively on material from my book *Reconstructing the Cognitive World: the Next Step* (Wheeler 2005), especially chapters 2, 7 and 10. Sometimes text is incorporated directly. Having said that, my re-use of that material here is not simply a re-hash of it. The present treatment has some new things to say about Descartes' enduring legacy in the science of mind.

[2] All quotations from, and page numbers for, Descartes' own writings are taken from the now-standard English editions of the texts in question. For the texts referred to here, this means the translations contained in (Cottingham et al. 1985a, 1985b).

[3] The first two of these notions are identified in Descartes' work by Hatfield (1992, 360-2). The third is not.

[4] For Descartes, the essential property of matter is that it takes up space, i.e., that it has extension. In effect, mechanics studies changes in manifestations of that property.

[5] For the view that useful fictions can be explanatorily powerful, see (one common way of understanding) Dennett's position on psychological states such as beliefs and desires (Dennett, 1987). Post-Darwin, the overwhelming temptation will be to see natural selection as the source of functional normativity in the case of the bodily machine. On this view, the function of some bodily element will be the contribution that that element has made to survival and reproduction in ancestral populations. Descartes, writing two hundred years before Darwin, didn't have this option in his conceptual tool-kit.

[6] For a more detailed description of these mechanisms, see (Hatfield 1992, 346).

[7] In the present context, the fact that AI came to mechanize general-purpose reason is plausibly interpreted as a move *against* Descartes. However, this is not the only way of looking at things. Aside from its mechanization, nothing about the nature and contribution of reason as a psychological capacity underwent significant transformation in the process of appropriation by AI. Thus, viewed from a broader perspective, one might argue that, by mechanizing general-purpose reason in the way that it did, AI remained within a generically Cartesian framework. For much more on this, see (Wheeler 2005, especially chapters 2 and 3).

[8] Roughly speaking, design by artificial evolution works as follows: First one sets up a way of encoding potential solutions to some problem as genotypes. Then, starting with a randomly generated population of potential solutions, and some evaluation task, one implements a selection cycle such that more successful solutions have a proportionally higher opportunity to contribute genetic material to subsequent generations, i.e., to be 'parents.' Genetic operators analogous to recombination and mutation in natural reproduction are applied to the parental genotypes to produce 'children,' and (typically) a number of existing members of the population are discarded so that the population size remains constant. Each solution in the resulting new population is then evaluated, and the process starts all over again. Over successive generations, better solutions are discovered. In GasNet research, the goal is to design a network capable of achieving some task, and artificial evolution is typically allowed to decide fundamental architectural features of that network, such as the number, directionality, and recurrency of the connections, the number of internal units, and the parameters controlling modulation and virtual gas diffusion.