Susan G. Sterrett Philosophy

# Too Many Instincts: Contrasting Philosophical Views on Intelligence in Humans and Non-Humans

Susan G. Sterrett
Wichita State University

# Too many instincts: contrasting philosophical views on intelligence in humans and non-humans

SUSAN G. STERRETT
*Department of Philosophy , Duke University, Durham, NC 27708, USA*
*e-mail: sterrett@duke.edu*

**Abstract**
This paper investigates the following proposal about machine intelligence: that behaviour in which a habitual response that would have been inappropriate in a certain unfamiliar situation is overridden and replaced by a more appropriate response be considered evidence of intelligence. The proposal was made in an earlier paper (Sterrett 2000) and arose from an analysis of a neglected test for intelligence hinted at in Turing's legendary 'Computing Machinery and Intelligence'; it was also argued there that it was a more principled test of machine intelligence than straightforward comparisons with human behaviour. The present paper first summarizes the previous claim then looks at writings about intelligence, or the lack of it, in animals and machines by various writers (Descartes, Hume, Darwin and James). It is then shown that, despite their considerable differences regarding fundamental things such as what kinds of creatures are intelligent and the relationship between reason, instinct and behaviour, all of these writers would regard behaviour that meets the proposed criterion as evidence of intelligence. Finally, some recent work in employing logic and reinforcement learning in conjunction with 'behaviour-based' principles in the design of intelligent agents is described; the significance for the prospect of machine intelligence according to the proposed criterion is discussed.

*Keywords*: Intelligence, animal mind, William James, Descartes, Hume, Darwin, Turing, instinct reason

*Received December 2001; accepted March 2002*

## 1. Introduction

'What kind of behaviour is evidence of intelligence?' When the question comes up, it is usually asked about machines or (non-human) animals. In this paper, I first recount a suggestion about what should count as evidence of intelligence that I made on an earlier occasion: roughly, the ability to override a habitual response and replace it with a more appropriate one (Sterrett 2000, 2002). That suggestion was inspired by a test described by Alan Turing, but the test that inspired it is underappreciated, as it was overshadowed by Turing's subsequent endorsement of another test for thinking that consisted merely in directly comparing the performances of a human and a machine in conversation. I argued that the suggestion inspired by the neglected version is especially appropriate for *machine* intelligence, since it does not depend upon the machine having any particular capacities.

After explaining that suggestion, I go on here to examine what a variety of thinkers-Rene Descartes, David Hume, Charles Darwin and William James-have had to say about the existence of intelligence in non-humans as well as in humans. I conclude that-despite the fact that these thinkers have strongly contrasting views on fundamental topics such as the relationship between instinct and reason, the relationship between mechanism and rationality, and the similarities and differences between humans and

animals-a case can be made that my suggestion characterizes a kind of behaviour that would count as evidence of intelligence on *all* of these thinkers' views.

There are two very different things one could take an answer to the question of what kind of behaviour would be evidence of intelligence to be saying:

> (i) there is some specific unobserved thing (e.g. thinking, consciousness, a soul) of which the intelligent behaviour is suppose d to be a sign; or
> (ii) whatever the unobserved process or entity giving rise to it, the behaviour in question would properly be considered intelligent.

The latter way of taking an answer to the question need not be a matter of *defining* intelligence, however. That is, one can take the question to be asking what would be good grounds for inferring intelligence while yet remaining noncommittal as to what is essential to, or what causes, intelligence.[1] We do this with behavioural concepts other than intelligence all the time; consider how we apply judgements such as 'kind' or 'stubborn'. The virtue of the second way of taking the question is that it is more open to the kinds of beings or things that could exhibit intelligent behaviour.

In the 17th century, Rene Descartes discussed how we would be able to differentiate a sophisticated human-like automaton from a genuine human, or a monkey-like automaton from a monkey. The question Descartes had in mind was clearly of the sort described as (i) above. Specifically, Descartes was addressing the question of whether naturally-occurring humans and monkeys consisted of anything more than the physiology he had researched so exhaustively. He thought the answer was different for humans and monkeys, and what he appealed to as a basis for this answer was the conclusion of another line of reasoning: although it was possible that there could be a monkey-like automaton that was behaviourally indistinguishable from a naturally-occurring monkey, any human-like automaton would necessarily be behaviourally distinguishable from a naturally-occurring human. There were two kinds of behaviour that were evidence of the existence of a rational soul, and so, he asserted with confidence, would not be exhibited by even the most sophisticated automaton. The first was that, while such an automaton might well be able to produce different words in response to different things being said or done to it, it would not be able to 'respond appropriately to whatever was said to it'. For Descartes, the ability to respond appropriately to whatever is said was something any human could do, and was evidence of the existence of a rational soul; reason, unlike any physiological organ, was a 'universal instrument'. The second was that, no matter how well an automaton might perform some tasks, it would necessarily flounder at some task or other at which a naturally-occurring human would be competent. Again, the crucial point was that reason is a universal instrument, and that it provides at least some help in producing responses to novel situations, as opposed to the kind of specialized help provided by instincts and habits.

Centuries later, in 1950, Alan Turing proposed a practical test to determine whether a machine might properly be regarded as intelligent. Actually, as I have pointed out in other papers (Sterrett 2000, 2002) there are two distinct tests in Turing's essay 'Computing machinery and intelligence' (Turing 1950). In my view, one has been unduly neglected, while the other has become extremely well-known. The well-known one, which I shall refer to as 'the standard Turing test', recalls the scenario Descartes imagined of being able to distinguish an automaton from a naturally-occurring human due to the automaton' s inadequacies in answering questions. In Turing's imagination, it is possible that things could turn out differently than Descartes imagined they must, however: Turing envisions the possibility that someone is

*not* able to easily distinguish the automaton from the human, based on their responses to what is said to them. However, I think it is not only the outcome of Turing's thought experiment that differs from Descartes' but also that the question Turing was addressing was different. Turing was addressing a question of the sort described in (ii) above, whereas Descartes was addressing a question of the sort described in (i) above. That is, the question for Turing was not whether the automaton's inner workings were the same as those of a naturally- occurring human's, but whether, to put it roughly, we would feel it appropriate to apply the word 'intelligent' to a machine if it behaved as the machine did in the scenario he imagined. Thus, the fact that Descartes and Turing draw different conclusions as to whether a machine could exhibit intelligent behaviour does not mean that Descartes and Turing would disagree as to what would count as intelligent behaviour.

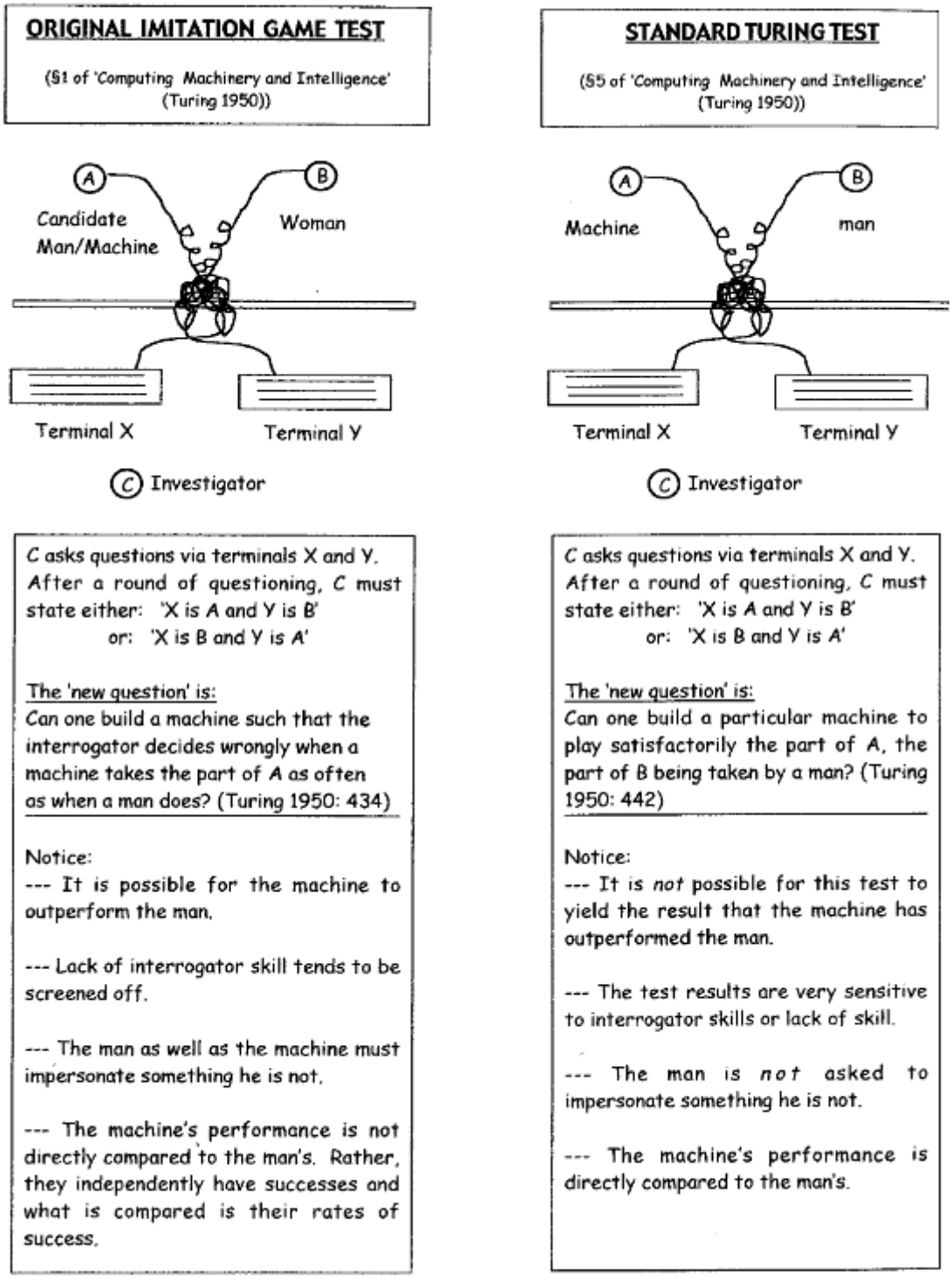
Others have noted the similarity between the well-known standard Turing test and what is sometimes referred to as 'Descartes' language test for animal mind', so this last point is hardly new. However, using the ability to converse as a criterion of intelligence is wrought with difficulties. Instead , I propose we look at the question of the possibility of machine intelligence using, instead of the standard Turing test, another test inspired by Turing: the test I refer to as 'the original imitation game test' (Sterrett 2000, 2002). We shall find something there that helps us with the question of how to tell machine intelligence if and when we see it.

**2. The importance of overriding habit**

Those who have read Turing's 1950 paper may recall that in it he proceeds roughly as follows. The paper starts out with the question 'Can machines think?' But, as he wants to avoid ambiguities due to the words 'machine' and 'think', Turing replaces the question 'Can machines think?' with a 'new question' that involves a parlour game. Then, he works out what he calls a 'more precise form' of this new question, introducing a simplified version that does not involve the parlour game. One could attempt to dig out a theme common to the two games Turing proposed and try to infer what he must have been after. However, in attempting to do this, I realized that the two tests would not always yield the same results; in fact, the results they yield are not even comparable. What I want to do in this section is explain how the first, neglected, test in Turing's 1950 paper-the test I call 'the original imitation game test'-works, and why I say that it is not equivalent to the second test, which is widely known as 'the Turing test'. I will also explain why I say that the first, neglected test employs a better characterization of intelligence. This section may be skipped by those familiar with the presentation in 'Turing's Two Tests for Intelligence' (Sterrett 2000).

The first test Turing proposed uses what is sometimes referred to as 'the original imitation game'. As this game is nested inside the test he described, I refer to the test itself as 'the original imitation game test'. This is the underappreciated and neglected test. It is depicted in the left-hand column of figure 1. In 'the original imitation game test', there are three players, each with a different goal: A is a man, B is a woman, and Cis an interrogator who may be of either gender. C is located in a room apart from A and B, and knows them by the labels X and Y. C asks questions of X and Y. The game is set up so as not to allow C any clues to 'X' and 'Y's identities other than the linguistic exchanges that occur within the game. In Figure 1, the teletypes are labelled X and Y; Turing also said, that, instead of teletypes, an intermediary could be used to communicate the answers from A and B to C.

Figure 1


ORIGINAL IMITATION GAME TEST
(§1 of 'Computing Machinery and Intelligence' (Turing 1950))
A
B
Candidate Man/Machine
Woman
Terminal X
Terminal Y
C Investigator
C asks questions via terminals X and Y. After a round of questioning, C must state either: 'X is A and Y is B' or: 'X is B and Y is A'
The 'new question' is:
Can one build a machine such that the interrogator decides wrongly when a machine takes the part of A as often as when a man does? (Turing 1950: 434)
Notice:
--- It is possible for the machine to outperform the man.
--- Lack of interrogator skill tends to be screened off.
--- The man as well as the machine must impersonate something he is not.
--- The machine's performance is not directly compared to the man's. Rather, they independently have successes and what is compared is their rates of success.
STANDARD TURING TEST
(§5 of 'Computing Machinery and Intelligence' (Turing 1950))
A
B
Machine
man
Terminal X
Terminal Y
C Investigator
C asks questions via terminals X and Y. After a round of questioning, C must state either: 'X is A and Y is B' or: 'X is B and Y is A'
The 'new question' is:
Can one build a particular machine to play satisfactorily the part of A, the part of B being taken by a man? (Turing 1950: 442)
Notice:
--- It is not possible for this test to yield the result that the machine has outperformed the man.
--- The test results are very sensitive to interrogator skills or lack of skill.
--- The man is not asked to impersonate something he is not.
--- The machine's performance is directly compared to the man's.

The game is played as follows: C interviews ‘X’ and ‘Y’ (as he knows them). At the end of the game, C is to make one of two statements :

> ‘ ‘‘X’’ is A (the man) and ‘‘Y’’ is B (the woman)’ or
> ‘ ‘‘X’’ is B (the woman) and ‘‘Y’’ is A (the man)’.

C’s goal is to make the correct identification, B’s goal is to help C make the right identification, and A’s goal is to try to fool C into making the wrong identification, i.e., to succeed in making C misidentify him as the woman.

The first formulation Turing proposed (in §1 of Turing 1950) as a substitute for ‘Can Machines Think?’ was this: ‘What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?’ Now, suppose we take this to be suggesting a kind of meta-game, of which the interrogator is unaware. That is, the interrogator C thinks that he or she is interviewing a pair consisting of one man and one woman. There are numerous rounds of such interviews, and at the end of the series of rounds, the frequency with which the interrogator misidentified the woman when the machine played the part of A is compared with the frequency with which the interrogator misidentified the woman when a man played the part of A. It is this meta-gam e that I call ‘the original imitation game test’.

In the subsequent discussion, Turing stated that, in turn, the question: ‘Are there imaginable digital computers which would do well in the imitation game?’ was equivalent to the following question:

> Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate program, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man? (Turing 1950: 434)

Turing is not explicit about what the interrogator is to determine in this second version of the game, but the standard reading is that the interrogator is to determine which player is the computer and which is the man. Such a reading seems plausible enough, as the interrogator’ s task would be parallel (though not identical) to the task in the first version of the game, i.e. at the end of the interview, the interrogator is to state one of two things: either ‘ ‘‘X’’ is A and ‘‘Y’’ is B’, or ‘ ‘‘X’’ is B and ‘‘Y’’ is A’, here, however, A is the compute r and B is the man.

The test for machine intelligence in this second version is then simply how difficult it is for an ‘average’ interrogator to correctly identify which is the computer and which is the man. This is what I refer to as the standard Turing test.

Figure 1 lists some major differences between ‘the original imitation game test’ and the standard Turing test. These differences may not be obvious on the surface, but they become obvious when we imagine running the test under a variety of circumstances. The first difference becomes apparent in considering the quantitative results the two tests can yield. In ‘the original imitation game test’, there is nothing in the structure of the game to prevent a machine scoring higher than a man: consider an outcome in which the man succeeds one percent of the time, and the machine succeeds three percent of the time. Here there is an independent measure of success other than the interrogator’ s direct comparisons of the responses of the man and the machine, because there is a task at which each can either fail or succeed, independently

of the other. If the interrogator wrongly identifies the candidate A as B when the man takes the candidate seat, the man has succeeded at the task at hand, and if the interrogator wrongly identifies the candidate A as B when the machine has played the role of candidate A, the machine has succeeded at the task at hand.

The standard Turing test does not even admit of such a result; there, the responses of the man and the machine are directly compared and only the machine is judged as to having exhibited a requisite skill or having failed to do so. The second difference is that the standard Turing test is far more sensitive to the interrogator's skill. That is, if C is very skilled at discerning the relevant differences, the percent of time the human wins will increase; if C is not very skilled at this, the percent of time the machine wins will increase. Sensitivity of the results of a test to the skill of the interrogator running it is hardly a virtue of any test. The results of 'the original imitation game test' are not nearly as sensitive to interrogator skill as the standard Turing test is, for in 'the original imitation game test' the machine's intelligence is being tested by comparing the frequency with which it succeeds in causing the interrogator to wrongly identify it as the woman with the frequency with which a man succeeds at doing the same. Thus, C's skill level affects both frequencies. More fundamentally, the differences between the two tests are due to this: In 'the original imitation game test', unlike in the standard Turing test, scoring as a thinker does not amount to simply being taken for one by a human judge.

Finally, there is a fundamental difference in what is asked of the man employed in the two tests. In 'the original imitation game test', both the man and the computer are called upon to impersonate. What the test results reflect is their relative success in achieving this goal. In the standard Turing test, the man is not called upon to do anything very novel, whereas the computer must converse under the pretence that it is human! In contrast, 'the original imitation game test' compares the abilities of man and machine to do something that requires resourcefulness of each of them. It seems to me very significant that, for both the man and the machine, the task set has these aspects: recognizing an inappropriate response, being able to override the habitual response, and being able to fabricate and replace it with an appropriate response.

Thus, I concluded:

> the importance of the first formulation lies in the characterization of intelligence it yields. If we reflect on how 'the original imitation game test' manages to succeed as an empirical, behaviour-based test that employs comparison with a human's linguistic performance in constructing a criterion for evaluation, yet does not make mere indistinguishability from a human's linguistic performance the criterion, we see it is because it takes a longer view of intelligence than linguistic competence. In short: that intelligence lies, not in the having of cognitive habits developed in learning to converse, but in the exercise of the intellectual powers required to recognize, evaluate, and, when called for, override them. (Sterrett 2000: 558)

This insight came from thinking about what is required of a man sitting in the candidate seat in 'the original imitation game test'. Carrying on a conversation employs lots of cognitive habits. But the task set in this test requires that he reflect on those habitual responses in case some of them ought to be edited to be the kind of behaviour expected of the one he is impersonating. Here it is not only content in his responses that may need to be edited, but tone (i.e. deference, concern, sarcasm about certain things and modesty about others) as well. But, rather than stating the criterion in terms of a game that only a creature with the ability to converse could participate in, I have tried to generalize the criterion by stating it in terms of the behaviour of a creature with instincts and habits, but not necessarily an ability to converse. Thus, on this criterion, no creature, human or non-human, is precluded from exhibiting behaviour that

would be considered good evidence of intelligence simply because it is unable to converse. But being able to provide this kind of evidence does require having something that plays the role of instincts and habits.

What are the analogues of instinct and habit for a machine? To give a very general, provisional answer, the responses a machine would have to inputs or environmental stimuli prior to any learning-or perhaps some subset of such responses-might be considered analogous to instincts. The responses or patterns of responses the machine develops after learning might be considered analogous to habits. What then would it mean for a machine to override its habits and replace a habitual response with one that is more appropriate for the situation? On this analogy, if we knew a certain machine's analogues of instincts and habits, and observe d the machine in a situation in which the responses arising from these analogues of instinct and habit were not appropriate, and further observe d that in that situation, the inappropriate response that would result from its instinctual and habitual reactions was replaced by a response that was better suited to the situation, it would make sense to regard the behaviour as intelligent.[2] And, we might properly refer to the machine itself as intelligent.

On this view, then, the question of whether a machine can produce behaviour that would provide evidence of intelligence turns on questions about its ability to override (its analogues of) instincts and habits, where habits are construed broadly enough to include cognitive habits.

**3. Diverging views on reason, habit and instinct in humans and non-humans**

Numerous people discussing Turing's 1950 paper have noted that Descartes appealed to the ability to converse as one means of distinguishing creatures with reason from creatures whose behaviour was caused by mere mechanism. As I have mentioned above, Descartes also gave another means of distinguishing creatures that were mere machines from creatures with reason. Since the skills of creatures that were mere machines could at best be due to instinct or learned habit, they would do the things for which they had specialized instincts extremely well, perhaps even better than a creature with reason could. However, these machines would be hopeless at the things for which they had neither specialized instincts nor a habit acquired through training or experience. For Descartes, there was no question that all mechanism could ever give rise to besides instincts were habits acquired through training or some interaction with its environment. There could be no such thing as the intelligence of *machines*. If there was a question about *animal* intelligence, it was whether or not animals were *merely* machines, and Descartes himself was confident that animals were machines and did not have minds. But, setting aside how different Descartes' question was from the question of whether machines could behave intelligently, let's consider what he would have made of the insight I believe is reflected in using frequency of success relative to a man in 'the original imitation game test'.

The criterion for intelligent behaviour I associate with success in 'the original imitation game test' does not rely upon having linguistic skills-rather, it is a more generalized criterion: whether a creature is able to recognize when its habits or instincts should be overridden, and is able to replace the instinctual or habitual response with one that is more suitable. So, we want to look at whether Descartes would have viewed such behaviour as evidence of intelligence. To put Descartes' view in perspective, it is helpful to look at later writers who knew Descartes' view as well. Darwin cited evidence for his claim that (contra Descartes) animals were *not* merely machines; is Darwin's reasoning in line with the proposed criterion?

What about other philosophers who investigated the workings of the mind and included non-human animals in their accounts, such as David Hume and William James?

In the chapter entitled 'Instinct', James describes animals that have lost what he calls 'the ''instinctive'' demeanour'. The animal that has lost the instinctive demeanour, James says, appears to lead 'a life of hesitation and choice, an intellectual life; *not, however, because he has no instincts-rather because he has so many that they block each other's path*' (James 1983: 1013). James was employing a commonsense and uncontroversial notion of instinct-roughly, that of a faculty of acting with neither foresight of the ends nor education in performing the act.[3] He wasn't challenging the notion of instinct, but he was challenging a prevailing view when he associated an intellectual life with having *too many* instincts. For he points out that 'Nothing is commoner than the remark that Man differs from lower creatures by the almost total absence of instincts, and the assumption of their work in him by ''reason''' (James 1983: 1010). James went on to define and categorize the kinds of instincts humans have, concluding that, in contrast to the commonly-held view: 'no other mammal, not even the monkey, shows so large an array [of instinctive tendencies as humans do]' (James 1983: 1056).

The view David Hume articulated falls under what James called the commonly-held view-the view that reason takes over the work of instincts. In *An Enquiry Concerning Human Understanding,* Hume associated reason with a paucity of specialized instincts-though not with a total absence of instinct. But, because he distinguished something he called experimental reason from what he called demonstrative reason, Hume's view is not quite so simple as the view James was challenging.

In comparing the abilities of humans and animals, Hume concluded that animals as well as humans learned from experience. He also recognized that animals have specialized instincts that produce behaviour not based on any such observation or experience: he cites the instincts birds have for incubation of eggs and nest building. Hume noted wryly that we humans are apt to admire such specialized instincts 'as something very extraordinary, and inexplicable by all the disquisitions of human understanding' (1977: 72). However, he regards this amazement wryly because he regards the ability to make inferences about matters of fact, whether it is found in humans or non-human animals, as itself an instinct. That is, he refers to the operations of the mind that give rise to beliefs and expectation s concerning matters of fact as 'a species of natural instincts' (1977: 30). Hume regarded this kind of inference as an operation of the mind, but not a case of the use of demonstrative reason; he said: 'Animals . . . are not guided in these inferences by reasoning' (1977: 70). He thought that deductions arrived at by means of demonstrative reason were often fallacious and usually too slow to be of much use in matters requiring timely action anyway. In Hume's eyes, since instincts are much more reliable than the application of reason, Nature exhibited wisdom in arranging for such functions of the mind to be carried out by means of 'some instinct or mechanical tendency' (1977: 37). Yet, for all his roguishness in extending the word 'instinct' to include the ability to make associations based on contiguity, Hume's view is still traditional in that such experimental reasoning fills in to pick up the slack left when specialized instincts do not determine action. This is just the traditional view that William James was challenging: that the relationship between instinct and reason was that reason fills in where instinct does not supply an automatic response.

Hume was challenging a commonly-held view, too, but a different one: the view that there is a difference in kind between animal and human reason. Hume laid out the difference between the reason of human and non-human animals as a difference in where they are located on a continuum, as opposed to the

commonly-held view that it is a difference in kind. Hume not only said that the so-called 'experimental reasoning' by which we infer effects from causes is itself a species of instinct, but he held that it is an instinct that humans share with animals. Now, it does not escape Hume's notice that there is a great difference between humans and animals in terms of their reasoning abilities. But he denies that the existence of even a large difference in their reasoning abilities shows that the mental abilities of humans and animals are not on a continuum. He points out the many various ways in which one human can differ from another, resulting in a great difference in reasoning ability between two humans (Hume 1977: 71, n. 36). Some are a matter of enhanced abilities, others of deficiencies. For example, one person might surpass another in the abilities of attention, memory, and observation, or one person might just be able to carry out a longer chain of reasoning than another. One person might have much more experience, enhancing his ability to think of and employ analogies. Some deficiencies he mentions are confusing one idea for another, or being biased by prejudices or 'from education or party'. All these factors contribute to the great differences in human understanding with which we are familiar. Thus, Hume argues, *just as* the fact that there are great variations between the reasoning abilities of *one human and another* does not show that some humans have reason and others do not, *so* the fact that there are great differences between the reasoning abilities of *humans and animals* does not show that humans have reason and animals do not.

Hume's view that we're all on the same continuum is certainly meant to conflict with Descartes' in that, on Descartes' view, there is a *qualitative* difference between human and non-human creatures and it is marked by a *definite boundary*. The qualitative difference Descartes meant to mark out is between beings with reason and those without reason: for him, all animals ('beasts') and mechanical artefacts fall cleanly on one side of it, and *all* humans on the other side of it. I have mentioned the reasoning Descartes gave for such a qualitative difference. In more detail, it goes like this: First, he referred to the vast amount of work he had done on the physiology of the body, showing that he was as aware as anyone of the capabilities of physiological processes as well as of machines. I cite here what he says to show the level of physiological detail of his philosophical project:

> And then I showed what structure the nerves and muscles of the human body must have in order to make the animal spirits inside them strong enough to move its limbs . . . I also indicated what changes must occur in the brain in order to cause waking, sleep and dreams; how light, sounds, smells, tastes, heat and the other qualities of external objects can imprint various ideas on the brain through the mediation of the senses; and how hunger, thirst, and the other internal passions can also send their ideas there. And I explained which part of the brain must be taken to be the 'common' sense, where these ideas are received; the memory, which preserves them; and the corporeal imagination, which can change them in various ways, form them into new ideas, and, by distributing the animal spirits to the muscles, make the parts of this body move in as many different ways as the parts of our bodies can move *without being guided by the will*, and in a manner which is just as appropriate to the objects of the senses and the internal passions. This will not seem at all strange to those who know how many kinds of automatons, or moving machines, the skill of man can construct with the use of very few parts, in comparison with the great multitude of bones, muscles, nerves, arteries, veins and all the other parts that are in the body of any animal. (Descartes DV1 DMT ap. 56 p. 139)

Then Descartes deduced the kinds of functions that such a machine could perform. He concluded that a monkey could perform all the functions we observe it to perform on the hypothesis that its motions were caused by mere mechanism. From this he claims to be able to discern that 'all of their motions could arise

solely from that principle which is corporeal and mechanical' and thus that 'in no way could we prove that there is a thinking soul in brutes' (quoted from Massey and Boyle 1999: 94, translation).

What about a human? Descartes imagines a really sophisticated, really fine piece of machinery that is physiologically indistinguishable from a human and can imitate 'the motions of our bodies for all practical purposes'. However, he says, the situation is different than it was for the imitation monkey. He thinks it is different because he thinks we would be able to tell that even this extremely sophisticated mechanism was not a human. The imagined automaton can do a lot:

> 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 what you want of it, if you touch it in another it cries out that you are hurting it, and so on). (DV1 DMT ap. 57 p. 140)

But Descartes said it would be possible to tell it was a machine and not a human:

> But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do. (DV1 DMT ap. 57 p. 140)

Thus, since he judges that it is *not* possible that humans could be mere mechanism, whereas it is possible that other animals, such as monkeys, could be mere mechanisms, this marks a qualitative difference between humans and (non-human) animals. Descartes is explicit that this deficiency in the ability to use language exhibited by machines and animals marks a fundamental difference from humans:

> This shows not merely that the beasts have less reason than men, but that they have no reason at all. For it patently requires very little reason to be able to speak; and since as much inequality can be observed among the animals of a given species as among human beings, and some animals are more easily trained than others, it would be incredible that a superior specimen of the monkey or parrot species should not be able to speak as well as the stupidest child-or at least as well as a child with a defective brain - if their souls were not completely different in nature from ours. (DV1 DMT ap. 59 p. 141)

Descartes seems aware that, strictly speaking, he has demonstrated only that we *cannot* prove that animals do as a matter of fact have reason, but we *can* prove that humans must; humans could not do what they do without it. I say he seems aware that he has only demonstrate d that he cannot prove that animals do have reason, rather than being able to prove that they do not, because he uses a probabilistic argument ('it would be incredible that. . .') for the conclusion that beasts have no reason at all. Massey and Boyle (1999: 97) also note this point, citing what Descartes writes in a letter to Henry More: 'But although I take it as demonstrate d that one cannot prove that there is any cognition in brutes, I do not think on this account that one can prove that there is none at all, since the human mind does not enter into their hearts'.[4]

Descartes seems to make up for the lack of certainty of such arguments somewhat by supplying several of them. In the other reason he gives for knowing that the automaton that looks and functions physiologically like a human is not a true human, there is what I take to be a recognition of the amazing specialized instincts animals have: even if it could do some things really well, it would flounder at others. We would know by these flounderings that they 'did not act through understanding but solely through the disposition of their organs'. Now, Descartes seems to assume that this sort of thing would be a giveaway

that something was not a true human, but not that something is not a true dog or a true monkey. Descartes seems to think that such floundering is to be expected of dogs, monkeys and other beasts; not for humans, though! He says that 'reason is a universal instrument which is of service in all situations'; the use of reason is supposed to contrast with the basis for movements found in automaton s and animals. The behaviour of automatons and beasts arises from dispositions of their organs or parts. Yet, Descartes is aware that the instincts of animals *can* be overridden; beasts can be trained:

> So when a dog sees a partridge, it is naturally inclined to run toward it; and when it hears a gun fired the noise naturally incites it to run away. But nevertheless setters are commonly trained so that the sight of a partridge makes them stop, and the noise they hear afterwards, when [the bird] is fired on, makes them run up to it. (Descartes 1989: 48)

The mechanism is the same as it is for humans, and can result from one single, unplanned experience:

> Indeed this habit can be acquired by a single action and does not require long practice. Thus, when we unexpectedly come upon something very foul in a dish we are eating with relish, our surprise may so change the disposition of our brain that we cannot afterwards look upon any such food without repulsion, whereas previously we ate it with pleasure. And the same may be observed in animals. (Descartes 1989: 48)

So there seems to be very little a dog could do to prove to Descartes that it was thinking. Obeying commands is not evidence of reason to Descartes, nor is avoiding hazardous things. Descartes thinks we could never know that the aversion to the hazardous thing was not based on a past bad experience the animal had with that thing, or something very similar to it; thus we could never regard such behaviour as evidence that the dog had used reason to infer that the hazardous thing was something to be avoided. Descartes' view is that not only could even the most amazing instinct be the result of mechanism, but so could obeying a command or avoiding a hazard that requires it to override an instinct! Yet, he holds that the human capacity to reason could *not* be a result of mechanism and, so, that there is a qualitative difference between non-humans and humans. On that point, Descartes' view is quite clearly directly opposed to Hume's.

To be fair to their views, however, perhaps it should be pointed out that there are some other points about the reason of animals on which Descartes and Hume are not as far apart as they are often portrayed. Although Hume thought that animals made inferences, he thought they only drew the kind of inferences that arose from experimental reason acquired by habit; he distinguished this kind of experimental reason from the kind that involves study, thought, and the recognition that something is an instance of a maxim. Descartes thought that what looked like animals making inferences could be described in terms of what we would now call learned responses, and that what animals did did not require reason. So, if we were to roughly align Hume's contras t between demonstrativ e reason and experimenta l reason with Descartes' contras t between reason and learned responses, we may find in fact that Descartes and Hume are not pushing diametrically opposed claims about reason and habit on every point.

Yet, if we extend the ground of agreement between them as far as we can, there still remains an important difference between them. Hume thought that beasts and humans were best understood as just being at different places on the same continuum, whereas Descartes thought that there was a difference in kind between the most clever, most trainable non-human animal and humans whom he ranked lowest in the ability to reason.

I have a suspicion about the root of this difference. The point of disagreement at the root of this difference between Hume and Descartes lies, I think, in the significance they attach to what they called demonstrative reason-for Descartes, it was the mark of a thinking soul, and hence of the utmost significance, while Hume made impudent remarks about philosophers who made such a big deal about demonstrative reason.

In fact, Hume devoted a long footnote (Hume 1977: 28-29, n. 20) to arguing that the long-standing distinction drawn between the two species of argument called 'reason' and 'experience' is erroneous or at least superficial. He illustrates the difference in the kind of inferences that one can draw by the following pair of examples:

> (i) from the history of a Tiberius or Nero, dreading 'a like tyranny , were our monarchs freed from the restraints of laws and senates';
> (ii) coming to have that same apprehension on the basis of 'the observation of any fraud or cruelty in private life' and the application of a little thought.

Hume does not deny the usefulness of the type of argument generally referred to as 'reason' (as opposed to 'experience'). The kind generally referred to as experience is limited in that it can only support an inference from an experience d event that is 'exactly and fully similar' to the event inferred. Certainly, he agrees, there is some value in being able to extend the kind of things we can establish by 'some process of thought, and some reflection on what we have observed, in order to distinguish its circumstances, and trace its consequences'-but he argues there that it is a delusion to think that there is *any* species of reasoning that is *totally* independent of experience. For, he says, 'If we examine those arguments which… are supposed to be the mere effects of reasoning and reflection, they will be found to terminate, at last, in some general principle or conclusion, to which we can assign no reason' (Hume 1977: 28-29, n. 20). I think that it is the difference in their attitudes towards how distinctive and significant demonstrative reason is that results in Descartes and Hume having diametrically opposed judgements as to whether there is a difference in kind between animals and humans. For, Hume really does recognize there is a difference in the kind of reasoning that animals and people can carry out, but he does not think it is important enough to regard human and animal understanding as fundamentally different. He seems to think of reason as a sort of aid in letting us expand the range of our understanding, by letting us go farther in the consequences we can draw. But that kind of difference is just the kind of differences we know exist among various humans. Any science, no matter how impressive, is at bottom not based upon a fundamentally different sort of understanding than the kind of understanding animals have of the world, which comes from experimental reasoning.

James (1983) introduced a different continuum: one of an ever- increasing quantity of instincts. And, he puts humans at the *high* end of this continuum; i.e. humans have more instincts than any other creature. For James, too, reason's role is to be a sort of aid to instinct, but it does not constitute a fundamentally different *kind* of force than instinct: 'there is no material antagonism between instinct and reason. Reason, per se, can inhibit no impulses; the only thing that can neutralize an impulse is an impulse the other way. Reason may, however, make an inference which will excite the imagination so as to set loose the impulse the other way' (1983: 1013). Here James is talking about behaviour rather than inferring consequences, but we see an approach in keeping with Humes': the *kind* of thing (acting in ways to bring about certain ends, expecting an event) going on is the same in humans and other animals, but the kind of reason

humans are capable of is an aid in extending the range of that capability. Thus humans are capable of a range unreachable by other animals, and philosophers have *mistakenly* taken this as a difference in kind, rather than range, of that capability. Hume sees himself as correcting the 'difference-in-kind' misconception with respect to human understanding, whereas James corrects a 'difference-in- kind' misconception with respect to human behaviour.

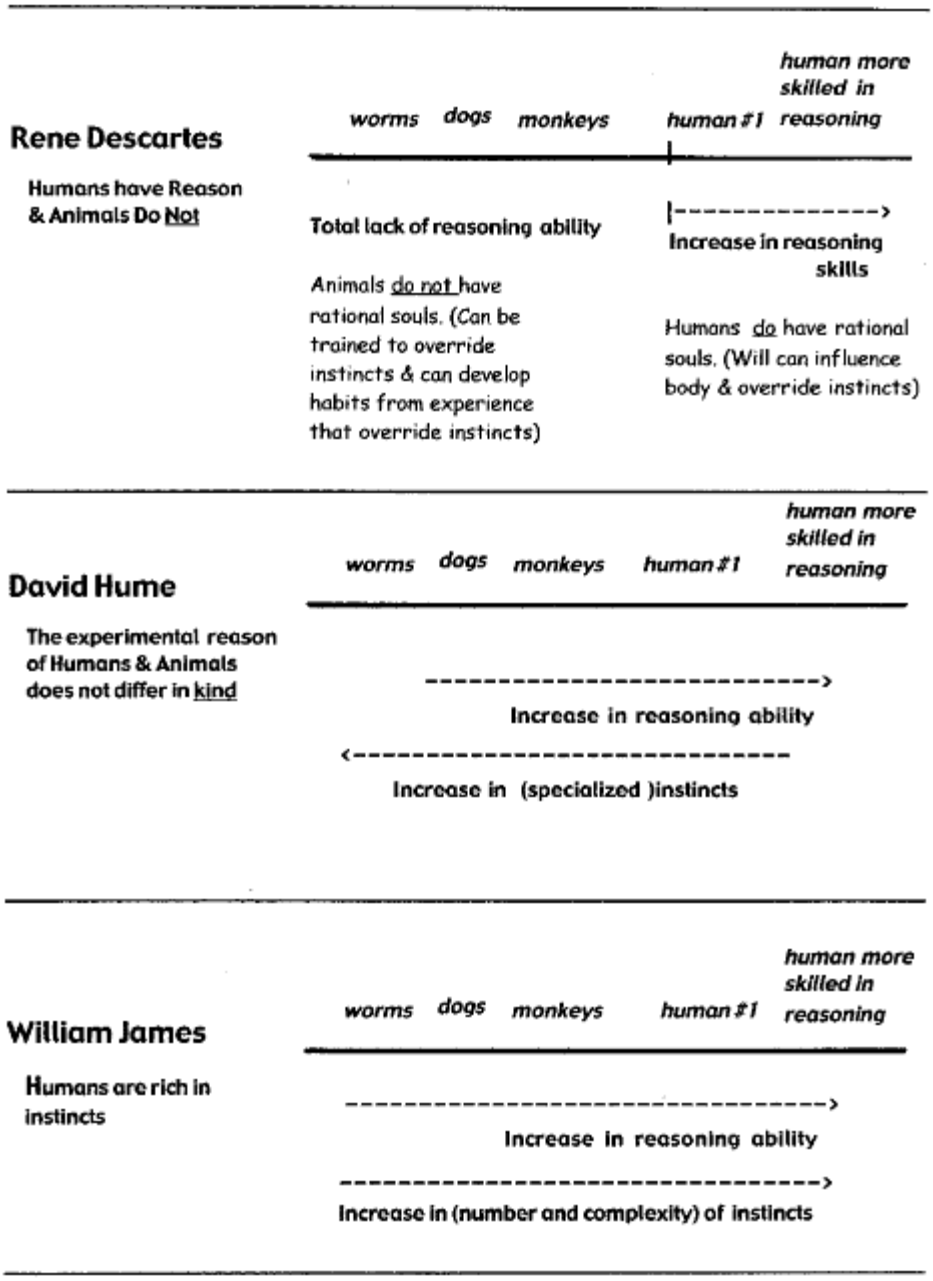
That is a brief sketch of how Rene Descartes in the 17th century, Hume in the 18th century, and William James in the 19th century, saw the roles of instinct, habit and reason in behaviour. Though they had different motivations and concerns, we can compare their views on these notions; these are depicted in figure 2. Now, let us ask how the suggestion I have made about what would be good evidence of intelligence fares on these various views.

**4. Converging views on looking smart**

Despite their very different views on the roles of instinct and reason in humans and animals, it turns out that these thinkers would give the same answer as to whether behaviour that meets the proposed criterion would count as intelligent behaviour.

We have already discussed Hume's view that understanding is a matter of degree, rather than an all-or-nothing matter. By looking at what he thinks is responsible for one person's understanding surpassing another's, we may be able to get some clues as to what Hume thinks counts as evidence of intelligence. A key factor Hume cited was the ability to form maxims based upon a particular observation or observations along with some thought and reflection.[5] Thus, intelligence would be exhibited by a creature's aversion to something that it had not ever experienced before on the basis of an inference employing a general maxim drawn from experience along with 'study and reflection'. The example he gave, cited in the last section, was of someone comparing a dread of tyranny based on knowing how terrible a particular tyranny such as Tiberius turned out, with a dread of tyranny based upon the observation of any fraud or cruelty in private life and 'the aid of a little thought' . In the latter case, there is not a close similarity between the event on which the dread is based and the thing dreaded. Thus, the ability to use thought to extend one's understanding of things to be dreaded and avoided beyond cases 'exactly and fully similar' to those known would be exhibited by behaviour in which, were the hazard or opportunity not present, the creature would behave in accordance with its habitual or instinctual response, but instead overrode that habit or instinct and avoided a hazard it is known not to have encountered before. Of course, as Descartes and Darwin noted, too, there are often problems ascertaining whether a response that appears thoughtful is not due just to instinct. I am here addressing the question of what would in principle be considered intelligent behaviour.

Figure 2




What about Descartes? We have some clue as to what Descartes thought would count as evidence of intelligence in animals because his argument proceeds by discussing the lack of such evidence. Descartes pointed out that animals can be trained to respond to specific words, and that some can even use words. But he thought that their responses to our verbal commands were limited to the sort that could be explained as a disposition of a particular organ, rather than requiring the use of the generalized instrument that reason is. The problem is, as he admits in a letter to Henry More, 'the human mind does not reach into their hearts' so we cannot prove that there is no cogitation in brutes. So even something that looks to

us like really intelligent behaviour can be accounted for on Descartes' view by appeal to the disposition of their organs either by design of nature or by acquired habit, including single-trial training. As I indicated earlier, it looks impossible to come up with something that would break through Descartes' blanket dismissal. However, Darwin finds some examples of cases that might count as evidence of cogitation in brutes that are not so easily dispensed with.

One example Darwin (1871) gave was of an experimenter who gave lumps of sugar wrapped in paper to his monkeys. Sometimes, Darwin (1871) says, the experimenter 'put a live wasp in the paper, so that in hastily unfolding it [the monkeys] got stung; after this had once happened, [the monkeys] always first held the packet to their ears to detect any movement within'. Notice that this response to the trauma of being stung is somewhat more sophisticate d than the simple aversion reaction Descartes mentioned: the monkeys instead initiate a new step in their sugar-eating ritual that will enable them to discriminate between the packets that will cause the same trauma and those that will yield only a pleasant tasty sugary experience.

But the example Darwin thought the best rejoinder to the kind of view Descartes held was of the animal for which he had the most affection and respect: a dog. He relates stories from two independent observers:

> *Mr. Colquhoun* winged two wild-ducks, which fell on the opposite side of a stream; his retriever tried to bring over both at once, but could not succeed; she then, though never before known to ruffle a feather, deliberately killed one, brought over the other, and returned for the dead bird. *Col. Hutchinson* relates that two partridges were shot at once, one being killed, the other wounded; the latter ran away, and was caught by the retriever, who on her return came across the dead bird; 'she stopped, evidently greatly puzzled, and after one or two trials, finding she could not take it up without permitting the escape of the winged bird, she considered a moment, then deliberately murdered it by giving it a severe crunch, and afterwards brought away both together. This was the only known instance of her ever having wilfully injured any game'. (Darwin 1871: ch. III)

Here it seems there was an unprecedented overriding of an instinct, not in response to some trauma, nor as a result of any training, but in order to achieve a desired end. It cannot be a disposition or an instinct, as these do not include foresight of the end. And this is why Darwin (1871) says he chose these examples: 'I give the above cases…because in both instances the retrievers, after deliberation, broke through a habit which is inherited by them (that of not killing the game retrieved), and because they show how strong their reasoning faculty must have been to overcome a fixed habit'.

It seems to me that Darwin is deliberately choosing cases that would meet Descartes' criterion, for he closes his discussion of anecdotes of monkeys and dogs that show the use of reason by saying that these exhibitions of reason by animals prove that animals are not machines. Thus, I think Darwin is employing the same criterion as Descartes, and I think he does so in order to challenge Descartes' conclusions. Just as Descartes was, Darwin is considering whether a creature known to possess instinct and habit might also use reason. Using reason is not a matter of a creature ridding itself of habit and instinct, but of the creature being able to exercise some power (which may be called reason) in recognizing, evaluating , and, when called for, overriding or perhaps modifying, the habitual response.

William James's bold and quirky step was to use the main insight in Darwin's challenging counterexample to Descartes, but to use it as an explanation of what gives rise to, rather than merely what

would give evidence for, the use of reason. James argues that having a plethora of instincts gives rise to behaviour exhibiting intelligence.

Now, I am not claiming James gives Darwin credit here; he does not. In general, James did not put much stock in anecdotes or 'dog stories' that purport to show that dogs reason. He discredits a story given him about a little terrier that is asked to retrieve a sponge and does so, in spite of never having been trained anything regarding the sponge. The little terrier's owner had said 'Sponge! Sponge! Go get the sponge!' anyway, while making motions of using a sponge to wipe up water, and was amazed and delighted at the dog's ability to discern what it was supposed to do. James gives the dog only a little bit of credit; he says:

> This terrier, in having picked those details out of the crude mass of his boat-experience distinctly enough to be reminded of them, was truly enough ahead of his peers on the line which leads to human reason. But his act was not yet an act of reasoning proper. It might fairly have been called so if, unable to find the sponge at the house, he had brought back a dipper or a mop instead. (James 1983: 974)

Now it seems that James is here actually calling for the kind of anecdote Darwin actually provided about bird-retrieving dogs, i.e. though he did not put it in such terms, what James is saying implies that what it would take to convince him that the dog used reason was a case where trying to carry out the habit of retrieving the object named would be frustrated and the animal would instead identify and carry out some substitute action that was appropriate to the situation.

James does not directly address Darwin's two example s of the problem solving bird retrievers, and, in fact, James (1983: 974) offers a universal dismissal of what he calls 'dog and elephant stories', stating with confidence that 'If the reader will take the trouble to analyse the best dog and elephant stories he knows, he will find that, in most cases, this simple contiguous calling up of one whole by another is quite sufficient to explain the phenomena'. James's point here about the sufficiency of 'calling up of one whole by another' to explain the animal's behaviour is that the behaviour is not evidence of the animal having to take the step of what he called 'dissociating characters', such as the character of being suitable for use in soaking up and removing liquid. Reasoning using characters would be evidence of using reason. James goes through numerous examples of 'dog stories' to discredit the claim that dogs can dissociate characters. There is one case he admits is borderline:

> Stories are told of dogs carrying coppers [pennies] to pastry-cooks to get buns, and it is said that a certain dog, if he gave two coppers, would never leave without two buns. This was probably mere contiguous association, but it is possible that the animal noticed the character of duality, and identified it as the same in the coin and the cake. If so, it is the maximum of canine abstract thinking. (James 1983: 975-976)

But James goes on to emphasize that most other stories of purported canine abstract thinking can be explained in terms of 'simple contiguous calling up of one whole by another'. Although James goes through one story after another, he never does get around to dealing with the stories Darwin cited to show that dogs were not mere machines. However, in spite of never mentioning the dogs Darwin wrote about who had to deal with competing instincts, James seems to have picked up on the relationship between having to deal with competing instincts and exhibiting intelligence. The kinds of things James lists among human instincts include the hunting instinct, the instinct to play in certain ways, curiosity, cleanliness and acquisitiveness, among numerous others. James thinks that having so many instincts is not only compatible with a more highly developed reason, but that having more instincts actually *leads to* a more highly developed reason. Here's why: the more instincts, wants and aesthetic feelings one has, the more

characters one will have to dissociate, and dissociating lots of distinct characters is a necessary part of making inferences that go beyond mere contiguity. Now, recall James's remark (James 1983: 1013) that 'there is no material antagonism between instinct and reason. Reason, per se, can inhibit no impulses; the only thing that can neutralize an impulse is an impulse the other way. Reason may, however, make an inference which will excite the imagination so as to set loose the impulse the other way', which I quoted earlier. James's discussion continues: 'and thus though the animal richest in reason might be also the animal richest in instinctive impulses, too, he would never seem the fatal automaton which a merely instinctive animal would be'. Thus, it seems that reason can be exhibited in cases where one is faced with competing instincts, and deals with it by employing reason to 'set loose' impulses that will neutralize others. Thus, although this is not exactly Descartes' view of reason overriding mechanical responses, the effect is the same: the realization of the appropriate behaviour called for does result in overriding a habitual or instinctive response that would be inappropriate.

It may be too strong to claim that there is some common notion or view held by all these thinkers in common. And some of them rule out the possibility of machine intelligence outright. My point is rather that, if we imagine asking the question: 'Given a creature possessed of habit and instincts that you do not rule incapable of intelligence outright in virtue of its constitution or genesis, what kind of behaviour would be evidence of its intelligence?' I think it can be argued that all four of these writers (though perhaps for very different reasons and with very different expectations about what the outcomes would be for various kinds of creatures) would consent that, were such a creature to show that it was capable of recognizing when the habitual or instinctual response was inappropriate, and be able to override it and replace it with an appropriate response, it would be evidence of intelligence.

**5. Intelligent robot designers**

I have deliberately attempted to steer the discussion about machine intelligence away from drawing comparisons between humans and machines, such as questions about whether machines could do some or all the things humans do or whether humans are machines. These questions are favourites of philosophers, and have often been connected with questions about the nature and existence of free will. I have instead here focused on what counts as evidence of intelligent behaviour in non-human s as well as in humans, and this question, interestingly, has led us to the importance of the capability of overriding instincts and habits.

The purpose of this investigation was to help us address questions about the nature and existence of machine intelligence. That the path has eventually led to the importance of the capability to override instincts and habits is doubly interesting, for research in robotic control and intelligent agents has come to the point of designing machine architectures that include not only implementation s of 'pre-wired' responses and reinforced learning, but of means of overriding and revising those 'pre-wired' and learned behaviours .

Although Turing seemed to think that the best way of constructing a simulation ofa human brain would be to build a machine that could learn, and would include a random element, the hearts of most of the first artificial intelligence program s were (deterministic) logical inference engines.[6] The learning involved, if any, was generally a matter of updating a representation of a domain by adding data received or inferred, rather than developing new procedures as a result of its interaction with its environment . Actions taken

were usually regarded as the result of carrying out a command produced by an inference engine that worked from goals to be achieved and, using information it had been given or had acquired, deduced the action or actions to take.

One of the most striking reactions against the centralized logical inference-engine conception of artificial intelligence reminds one of how Hume de-emphasized demonstrative reason and emphasized habit and instinct: I refer here to Rodney Brooks's advocacy of a 'behaviour-based' approach to robotic navigational problems. To avoid any misunderstanding, I should mention that Brooks has more recently tackled the task of developing humanoid robots, in which he explicitly looks to human beings as models for a humanoid robot design. I am interested in the question of machine intelligence, rather than in the specific question of how to build a human-like robot, and for this purpose I want to refer only to the basic change in orientation in AI he initiated in the 1980s, which continues to influence robotics today. He describes his behaviour-based approa h (circa 1991) thus:

> At MIT we have been investigating structuring of intelligence based on a decomposition into behaviours, each of which connects sensing to action. Functional modules such as planners and learners do not appear as such, but instead planning, learning, etc., can be observed to be happening in the complete system. The behaviours are the building blocks, and the functionality is emergent. This differs from the traditional approach in which the functional modules are the building blocks and the behaviours are emergent. (Brooks 1991)

The difference between Brooks' view and the traditional view he contrasts it with that is of interest to us here, however, is this: on Brooks' view, there is the possibility that, within the same machine or robot, different parts of the machine can call for different actuator responses of the same actuator, and hence can conflict. Thus the need for what Darwin (1871) called an 'inward monitor' can arise.

Brooks appreciated that his approach raised the problem of making decisions as to how to resolve conflicts between behaviours . Resolving conflicts was just one of two parts of the general problem of 'deciding which behaviour or behaviours should be active at any particular time', he said; the other was how to select potentially correct behaviours in the circumstances. The approach Brooks calls 'behaviour-based programming', is a refinement of his earlier 'subsumption architecture' approach. A distinctive feature of both these approaches is the use of layers each of which employs finite state machines (augmente d with timers), in contrast to a centralized decision-making unit. The layers are layers of behaviours , such as 'locomote', 'avoid hitting things' and 'explore', and the layers interact with each other. I do not wish to go into more detail about Brooks' specific methods here, as expositions of his approach are readily available elsewhere, but I think it worth remarking that his approach to implementing behaviour includes not only processes that can be halted, but processes that can be inhibited or blocked by others.[7] Thus Brooks' approach to intelligent behaviour recognizes, at least in principle, the need for the ability to determine an appropriate behaviour and for the ability to override and/or inhibit default processes. Brooks describes the layers as being related but thinks it improper to think of them as related by a hierarchy, as higher-level layers both depend upon and influence lower layers.

I have highlighted the features in his approach that relate to my claim about machine intelligence: the capabilities of determining appropriate behaviours in a situation and of being able to override instincts and habits. Brooks, however, emphasizes other features of this approach, especially the lack of symbolic logic processing in his approach. He notes: 'Neither [behaviour-based programming], nor the other

similar architectures, use symbols in the conventional way' (Brooks 1991b). He wants to highlight how his 'behaviour-based ' approach differs from more classical AI approaches , so he also makes the points that, on his approach , there is no centralized control unit, and no centralized representation of the world stored in the agent to be manipulated and employed in determining its next move.

However, other researchers who see the virtue of a behaviour-base d approach but do not see the virtue in doing without symbolic processing for the sake of it have employed architectures inspired by Brooks' layered 'subsumption architecture' in conjunction with the use of symbolic logic. For instance, in a paper entitled 'Logic-Based Subsumption Architecture', Amir and Maynard-Reid took on the task of extending a Brooks-style architecture to agents that are not necessarily physical, and to more complex tasks. They use logic rather than hardware to provide the link between a (behavioural) layer's inputs and outputs, but retain the features of Brooks' approach that are responsible for the reactivity of his robots. Corresponding to each of what Brooks called a layer, their architecture provide s a first-order logic theory, and 'Each layer is supplied with a separate theorem prover, allowing the layers to operate concurrently '(Amir and Maynard-Reid II 1999). They explain reactivity of the robot as an advantage arising from the layer-decoupling:

> As in Brooks' system, lower layers controlling basic behaviours are trusted to be autonomous and do not need to wait on results from higher layers (they assume some of them by default) before being able to respond to situations. Because these layers typically have simpler axiomatizations, and given the default assumptions, the cycle time to compute their outputs can be shorter than that of the more complex layers. (Amir and Maynard-Reid II 1999)

In keeping with Brooks' approach, however, the layers are not completely independent. It is important that the layers be able to interact. The possibility for 'conflict' between layers exists and hence so does the need for some approach to resolve the conflict. The connection with animal examples is that the layers are like habits or instincts, and intelligence (or lack of it) is exhibited in how the animal responds to conflicting instincts. Amir and Maynard-Reid II present one of several possible means of resolving the conflict of theories of various layers of their 'logic-based subsumption architecture' in detail, which they have successfully implemented in a mobile robot. The details are not important to my point in this paper. What is significant to me about their approach is that they implemented logic in a layered architecture in which logical processes take place not only at many different behavioural layers, but in dealing with conflicts arising from the interaction of layers as well.

In drawing an analogy between an animal and a robot using their architecture, I suppose an instinct would be analogous to something like a behavioural layer, so that the conflicts that arise between layers would be analogous to conflicting instincts. Thus, it is interesting that the problem of instincts conflicting and overriding each other arises in a robot using logical theorem-proving techniques throughout its operation.

However, the implementation reported in that paper is a navigational robot and, although they plan to extend its capabilities to making maps and reasoning about the world, they do not stress learning new skills. We are especially interested in learning, because many of the habits that figure in the exhibitions of intelligence we considered involve habits which, though they may get their start as instincts, involve things that are learned, and, in particular, learned in interactions with, or by observing, people. Here I have in mind anecdotes such as the little terrier figuring out what was being asked of him either from observation of his owner using the word 'sponge' on other occasions or his owner's use of gestures

indicating the use of a sponge to wipe up liquid while ordering him to get the sponge; the monkeys developing the habit of holding packets of sugar to their ears to listen for wasps occasionally put inside by their mischievous human keepers; the bird dogs sizing up a situation in which their default habits would not produce the usual result and coming up with a way to satisfy their owner's expectations by overriding one of those habits, and so on.

However, there is no reason that an architecture that has the features I find important to machine intelligence cannot also be much more flexible with respect to learning. Parr and Russell (1997) employed an approach to reinforcement learning, which they describe as 'providing a link between reinforcement learning and 'behaviour-based' or 'teleo-reactive 'approaches to control'. The emphasis of their presentatio n is on showing how reinforcement learning algorithms can benefit from the constraints imposed by a hierarchy of machines, as those constraints incorporat e prior knowledge to reduce the complexity of the problem. However, we are interested here in how the use of reinforcement learning enriches the abilities of a 'behaviour-based ' robot and makes it more flexible. Their approach uses a layered hierarchy of partially specified machines, in which the machines are non-deterministic finite state machines, and a higher-level machine may depend upon lower-level machines in making the transition from one state to another. The general approach permits a range of specification, but, 'One useful intermediate point is the specification of just the general organization of behaviour into a layered hierarchy, leaving it up to the learning algorithm to discover exactly which lower-level activities should be invoked by higher levels at each point'. The significance of this design to our investigation of machine intelligence is that, as I said earlier, one feature of the kind of behaviour that would count as evidence of intelligence is being able not only to override an inappropriate response, but to construct an appropriate one, and this approach seems to provide an architecture in which new solutions to problems are sought using an robot architecture that also contains analogues of instincts and habits.

This is by no means a survey of the field. These examples were selected to illustrate that as the design of machine architecture s progresses, we see that artificial intelligence researchers are designing algorithms and architectures that have the beginnings of some of the analogues of the features of behaviour that I identified at the outset of this paper as features of behaviour that would count as evidence of intelligence.

**6. Conclusion**

Let us return to the question at the start of my essay. I began by recapitulating the conclusion of an earlier analysis of the two descriptions of tests proposed by Alan Turing to replace the question, 'Can machines think?'. I had concluded that the two descriptions, taken literally, really describe two distinct tests, and that the test that has been neglected, though it seems at first rather eccentric and to test only a particular and rather peculiar skill, actually reflects an important insight. That insight is that behaviour that requires an agent to recognize when the response that would be produced by habit (including cognitive habits) is inappropriate, to override response in its place, would be considered intelligent behaviour.

In surveying various thinkers, we see that Descartes would have granted as much, although he would have thought only humans capable of such behaviour. Hume thought highly of habit, but even he too granted that there was a difference between the kind of reasoning due to habit in which one infers like from like, as opposed to the kind of reasoning in which one goes beyond habit. The difference is that he thought much more of our reasoning is due to habit than was generally recognized, and he did not think the

distinction was of nearly the significance that Descartes did. Darwin thought animals could exhibit the kind of behaviour I claim would be evidence of intelligence, and cited anecdotes of such behaviour as evidence that animals were not machines. So Darwin would agree that such behaviour was evidence of intelligence, but not that machines were capable of such behaviour. William James's view that reason arises from a plethora of instincts seems to indicate that he thought being able to arbitrate between them and override some by others was a sign of intelligence. I think the fact that there would be agreemen t as to what would count as evidence of intelligence despite such differing views on the topics of instinct and intelligence in humans, animals and machines is a point in favour of my claim, for it indicates that my claim is not tied to any particular intuitions about whether various sorts of non-human s are or are not intelligent. These thinkers all predate d Turing's famous paper of 1950 in which he suggested tests to replace the question 'Can machines think?'

Finally, looking forward from Turing's paper, I have given examples that indicate that as designers of intelligent agents and robots try to design architecture s that are more efficient and more flexible with respect to the range of situations that might be encountered in an environment , they are tending to come up with designs that have analogues of some of the feature s and abilities that I have said are ingredients of the kind of behaviour that would be considered evidence of machine intelligence. Though they have built-in analogues of instincts and the ability to develop habits, they also have, at least in principle, mechanisms with the potential to construct responses to replace the habitual or instinctual responses.[8]

Of course this does not say much about how a machine could autonomously determine when a habitual response is not appropriate or what an appropriate response to replace it would be. I am not claiming anyone has achieved this so far, nor that it is clear just how to achieve it. To the question of how we will agree as to when a machine's behaviour is not just a habitual response, but involved recognizing that a habitual or instinctual response was inappropriate and constructing an appropriate response to carry out in its place, though, I can say this: when you stop to think about it, the cross-gendering task set for a human, whose success the machine must match in order to pass the neglected test in Turing's essay, 'Computing Machinery and Intelligence', is just such a discriminating task.

## Notes

**1.** This point is not new and was made as early as 1976 by James Moor (1976).
**2.** Of course, what is 'appropriate' in a situation may depend in part on the role of a participant in it, so an adequate description of a situation might have to include something about analogues of goals and/or desires of the machine. Normally, the notion of an end will be obliquely built into the notion of instinct (see section 3 of this paper), but some situations might involve more detail (e.g. a dog's desire to carry out its master's command is an attachment to a specific person, whereas the instinct to attach itself to some human or other in a master-pet relationship is general and does not involve any particular person).
**3.** 'Instinct is usually defined as the faculty of acting in such a way as to produce certain ends, without foresight of the ends, and without previous education in the performance'. This is the opening statement of chapter XXIV 'Instinct', in *The Principles of Psychology* (James 1983: 1004).

**4.** I read Massey and Boyle's 'Descartes' Tests for Animal Mind' (1999) after I had written 'Turing's Two Tests for Intelligence' (Sterrett 2000). Their interesting discussion of Descartes' test for volitions (the 'action test') as evidence of animal mind stimulated me to think more about the rarely asked question of what, besides the ability to

converse, Descartes would have to count as evidence of animal mind. I highly recommend their innovative paper, which has led me to write this one connecting issues in history of philosophy about machines and intelligence with current ones in artificial intelligence.

**5.** 'There is no man so young and inexperienced, as not to have formed, from observation, many general and just maxims concerning human affairs and the conduct of life; but it must be confessed, that, when a man comes to put these into practice, he will be extremely liable to error, till time and farther experience both enlarge these maxims, and teach him their proper use and application' (Hume 1997: 29, n. 20)

**6.** In 'Computing Machinery and Intelligence', Turing describes what he calls a 'child-machine' in Section 7 of that paper (1950: 454). He says that the child-machine should have some analogue of pleasure and pain so that it could be taught using reinforcement learning, and remarks that the machine's behaviour should not be completely determined by its experience.

**7.** See entries for Brooks in the references section of this paper. As of the time of this writing, Brooks has an extensive webpage about his publications. (http://www.ai.mit.edu/people/brooks/publications.shtml).

**8.** Of course, it cannot be ruled out that the behaviour of a robot with such an architecture could be produced by one without such analogues of instincts and habits, i.e. someone could claim that a connectionist machine with just the right number of layers, just the right learning algorithms, and just the right kind of training set might be able to produce the same outputs. As long as the equivalence of behaviour is not sensitive to the particular characteristics of the novel situation in which the machine exhibits intelligence, I do not see why it would necessarily be a mistake to credit such a machine with intelligent behaviour either. (Were the machine's behaviour extremely sensitive to getting exactly the right kind of training, however, then I think some would want to withhold giving it credit, for the same reason that Descartes did so for highly trained responses of animals: the dependence on a very particular training is characteristic of a habitual response, and would raise suspicions as to how flexible the behaviour really was.)

**References**

Amir, E., and Maynard-Reid II, P., 1999, Logic-based subsumption architecture. 16th International Joint Conference on Artificial Intelligence (IJCAI'99).

Brooks, R. A., 1986, A robust layered control system for a mobile robot. *IEEE Journal of Robotic s and Automation*, 2: 14-23.

Brooks, R. A., 1990, Challenges for complete creature architectures. First International Conference on Simulation of Adaptive Behavior, Paris, France, September, pp. 434-443.

Brooks, R. A., 1991a , Intelligence without representation. *Artificial Intelligence Journal*, 47: 139-159.

Brooks, R. A., 1991b , Integrated systems based on behaviors. *SIGART Bulletin*, 2: 46-50.

Brooks, R. A., 1991c, Intelligence without reason. *Proceedings of 12th Int. Joint Conf. on Artificial Intelligence*, Sydney, Australia, August, pp. 569-595.

Darwin, C., 1859, *On the origin of species by means of natural selection* (London: John Murray).

Darwin, C., 1871, *The descent of man, and selection in relation to sex*, 2 vols (London: John Murray).

Descartes, R., 1985 *The Philosophical Writings of Descartes*, trans. John Cottingham, R. Stoothoff and D. Murdoch, 2 vols (Cambridge: Cambridge University Press).

Descartes, R., 1989, *The Passions of the Soul*, trans. Stephen H. Voss (Indianapolis, IN: Hackett Publishing Company).

Hume, D., 1977, *An Enquiry Concerning Human Understanding*, Eric Steinberg (ed.) (Indianapolis, IN: Hackett Publishing Company).

James, W., 1983, *Principles of Psychology* (Cambridge, MA: Harvard University Press).

Massey, G. J., and Boyle, D. A., 1999, Descartes's tests for (animal) mind. *Philosophical* Topics, 27: 87-145.

Moor, J. H. 1976. An analysis of the Turing test. *Philosophical Studies*, 30: 249-257.

Parr, R., and Russell, S., 1997, Reinforcement learning with hierarchies of machines. NIPS 97. Available online: http://www.cs.duke.edu/¹parr/ham-nips97.ps.gz

Sterrett, S. G., 2000, Turing's two tests for intelligence. *Minds and Machines*, 10: 541-559.

Sterrett, S. G., 2002, Nested algorithms and 'The Original Imitation Game Test': a reply to James Moor. *Minds and Machines*, 12: 131-136.

Turing, A. M., 1950, Computing machinery and intelligence. *Mind*, 59: 443-460.