AUTHOR QUERY FORM

ELSEVIER	**Journal: SHPS** **Article Number: 897**	**Please e-mail or fax your responses and any corrections to:** **E-mail: corrections.eseo@elsevier.sps.co.in** **Fax: +31 2048 52799**

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

Any queries or remarks that have arisen during the processing of your manuscript are listed below and highlighted by flags in the proof. Click on the 'Q' link to go to the location in the proof.

Location in article	Query / Remark: click on the Q link to go Please insert your reply or correction at the corresponding line in the proof
	We have frequently found errors in the citation to *Studies in History and Philosophy of Science, Studies in History and Philosophy of Modern Physics and Studies in History and Philosophy of Biological & Biomedical Sciences* and therefore would be grateful if you could confirm that the article you are citing is correctly referenced to the right journal

Thank you for your assistance.

Highlights

▶ I first show that the Turing test is not an expression of behaviourism. ▶ To demonstrate this, I outline Turing's necessary condition for intelligence. ▶ Then I show that Alan Turing was likely aware of Descartes's 'language test'. ▶ Last I argue that Descartes's and Turing's tests have similar epistemic purposes.

Studies in History and Philosophy of Science xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Descartes' influence on Turing

Darren Abramson

Department of Philosophy, Dalhousie University, Halifax, Nova Scotia, Canada

1. Introduction

Alan Turing, in his 1950 *Mind* paper 'Computing Machinery and Intelligence,' introduces what is now called 'The Turing test' (Turing, 1950). Turing's paper has inspired countless papers in support of, and critical of, the claim that computers can think. The received view of Turing's philosophy of mind is that he was a behaviorist. This view has persisted despite numerous critical evaluations of his work that argue to the contrary.

In this paper I begin by briefly comparing reasons that have been offered for the claim that Turing was not a behaviorist (despite his apparent commitment to the claim that thinking requires nothing more than displaying verbal behavior indistinguishable from a person). The strongest reason for understanding Turing this way, I argue, is his commitment to a non-behavioral necessary condition for intelligence. Then I show that

1. Turing was aware of Descartes' 'language test', and likely had it in mind when writing his 1950 *Mind* paper that introduces the Turing test; and,
2. Turing intended the imitation game to play an epistemological role that is similar to the role that Descartes intended the language test to play.

If Turing wasn't offering a behaviorist view, unlike many of his contemporaries, what non-behaviorist influences (if any) planted the seed in Turing's mind of what may seem, at first glance, a behaviorist understanding of thinking? I answer this question by a close reading of some of Turing's personal papers from the years immediately preceding the publication of the paper introducing the Turing test. With historical influences in place, I argue that, far from being coincidentally similar, Descartes' language test and Turing's imitation game are both intended as nearly certain tests for thinking, and as tests for internal, particular causes of thinking (although Turing and Descartes disagree on what the necessary internal causes of thinking are).

2. Turing and behaviorism

2.1. Definitions

In his 1950 article, Turing explains a party game he calls the 'imitation game.' In it, an interrogator (C) judge must determine, solely through written queries and responses, which of two other participants is a man (A) and which is a woman (B). The judge is aware that one of the participants is a man and one is a woman. Turing proposes to replace the question 'can machines think' with the following:

> 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? These questions replace our original, 'Can machines think?' (Turing, 1950, p. 434)

Let us fix our attention on one particular digital computer *C*. Is it true that by modifying this computer to have adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, *C* can be made to play satisfactorily the part of A [the man contestant] in the imitation game, the part of B [the woman contestant in the imitation game] being taken by a man? (Turing, 1950, p. 442)

Later commentators have, almost universally, interpreted the 'modified imitation game,' now called the Turing test, as follows: can a judge, communicating entirely through typed text, distinguish a human from a computer? This interpretation irons out two ambiguities of Turing's presentation. Some readers have argued that Turing intended the judge in the computer version of the imitation game to be answering a question about the gender of the players.[1] However, there is ample evidence that Turing did not intend the computer version of his test to involve gender issues.[2] Also, there is the question of what adequate performance amounts to. I will use the formulation that the judge distinguishes the computer from the person at a rate no better than chance.

E-mail address: da@dal.ca

[1] For example, Sterrett (2000).

[2] A compelling case, with an overwhelming (but not exhaustive) amount of textual support for the standard interpretation can be found in Piccinini (2000).

0039-3681/$ - see front matter
doi:10.1016/j.shpsa.2011.09.004

Given Turing's position, and the influential logical and methodological behaviorists he was contemporary with (Gilbert Ryle's *The Concept of Mind* had been published the previous year; B.F. Skinner's *Science and Human Behavior* would be published in 1953), a behaviorist interpretation of Turing's views is almost irresistible. On both sides of the Atlantic Ocean, there were pushes to understand the mind in terms of behavior. The Turing test, at first blush, is a paradigm of behaviorism: Turing says outright that the question of whether machines can think is strictly *meaningless*, and must be 'replaced' with the question of whether a machine can pass the imitation game (Turing, 1950, p. 442).

By the mid-1960s, the behaviorist interpretation of Turing's article was presented without critical assessment in popular philosophical texts. For example, consider the anthology *Minds and Machines*, edited by Alan Ross Anderson (1964), which contains as its first article Turing's 'Computing Machinery and Intelligence'. The second article reprints Michael Scriven's 1953 article, also published in *Mind*, but with a short addendum. The first part of the addendum reads: "This article can be taken as a statement of the difficulties in (and was written partly as a reaction to) Turing's extention [sic] of behaviorism into the computer field." (Scriven, 1964, p. 42). *Minds and Machines* collected a number of the most significant articles of the era on the computational theory of mind by Turing, Scriven, Keith Gunderson, J.R. Lucas, and Hilary Putnam. I am not claiming that this anthology engendered the behaviorist interpretation of Turing, but it is an early sign of its widespread adoption.

2.2. A first response: Turing's consideration of 'contrary views'

As mentioned above, Turing uses language that unequivocally asks for the replacement of questions of machine mentality with questions of behavior, and does so by appealing to the meaninglessness of the former questions. The replacement of mentality questions with behavioral questions appears to betray a commitment to something like a verificationist criterion of meaning. A verificationist interpretation of the Turing test, though, is inconsistent with much of what Turing says in his article.

As some have pointed out,[3] the sixth section of Turing's 1950 paper deals with objections that he can't seriously consider if he takes the passing of his test to be *equivalent* to the possession of mind. I will mention just two examples. In his consideration of the mathematical objection, Turing takes seriously the idea that there might be in-principle limitations that distinguish computers from humans, and that one might not be able to ascertain whether the candidates in the imitation game were subject to those limitations. Turing's response does not at all address the *detectability* of those limitations in the test, but in fact *denies* that computers are subject to them while humans are not.[4] In his consideration of the argument from consciousness, Turing considers at length an objector who claims that despite excellent performance in the test, a machine wouldn't have consciousness. Turing *does not* respond by denying the reality of inner conscious states to mental entities. Instead, Turing offers the skeptic a parity argument, according to which consciousness can no more be denied of machines that pass the test than it can be denied of other people.[5]

Still, someone might argue, this only shows that Turing lacks consistency in his presentation; although he veers away from his behaviorist line in defending from criticism the claim that machines can think, his central positive project retains a criterion for thought that is behaviorist. This claim can be answered: there are stronger reasons for denying Turing a behaviorist interpretation, which I will now present.

2.3. The second response: necessary vs. sufficient conditions

Many commentators have pointed out that 'Computing Machinery and Intelligence' *cannot* be read as a bare statement of logical behaviorism.[6] Here is the passage that obviates such an interpretation of the article:

> May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection (Turing, 1950, p. 435).

In this passage, Turing reveal that he is only committed to the sufficiency of passing the test for thinking, and not its necessity. Therefore, he cannot be offering a *definition* or an *analysis* of thinking. Put more simply, the passage quoted is consistent with the existence of thinking things that don't display the particular behaviors under consideration. Logical behaviorism purports to provide the *referents* for mental terms—so, Turing cannot be a logical behaviorist.[7]

Many have noticed this property of Turing's position (its 'sufficiency behaviorism') and object to this, calling it behaviorism all the same. For example, Ned Block targets this view of Turing's directly, and views prior attacks on behaviorism as deficient *because* they don't rule out sufficiency behaviorism (Block, 1981, pp. 15–16). John Searle also identifies sufficiency behaviorism and makes it his target: "The Turing test, as you will have noticed, expresses a kind of behaviorism. It says that the behavioral test is conclusive for the presence of mental states" (Searle, 2004, p. 70).

Once Turing's position has been thus clarified, many have been happy to simply call the view that behavior is sufficient for intelligence a form of behaviorism, thus reuniting Turing's views in a general way with his famous contemporaries. There are many well known criticisms of Turing's claim that passing the Turing test is sufficient for thinking[8], but I will not wade into these debates. Instead, I will now show that Turing is not, in fact, a strict sufficiency behaviorist.

2.4. The third response: the strength of the Turing test

A third, subtle response one could make to the charge of behaviorism is that Turing is committed to the view that his test only provides a sufficient condition for intelligence *because it measures some non-behaviorally defined property*. Now, if Turing understands his own test this way (as I will argue he does), then whether or not one *agrees* that the test is a measurement tool for this non-behavioral property, Turing is not even a strict 'sufficiency behaviorist'. That is, Turing cannot be understood as believing that possessing certain behaviors is always sufficient for intelligence. Instead, the interpretation goes, possessing certain behaviors is *evidence* for some other property, and the possession of the other property is required for intelligence.

This interpretation of Turing is not new. It is offered first by James Moor, who describes success in the Turing test as 'inductive

[3] For example, Leiber (1995, p. 63).
[4] See, for example, Turing (1950, pp. 444–445).
[5] I am not, of course, endorsing Turing's response to a skeptic of computer consciousness, but merely pointing out the inconsistency of Turing's response with behaviorism.
[6] For example, see Block (1981, pp. 15–16).
[7] Daniel Dennett makes this point. See Dennett (1985, p. 4).
[8] The most famous criticisms are by Ned Block and John Searle, namely the Chinese gym and Chinese room arguments, respectively.

Please cite this article in press as: Abramson, D. Descartes' influence on Turing. *Studies in History and Philosophy of Science* (2011), doi:10.1016/j.shpsa.2011.09.004

evidence' of thinking, where the conclusion that something thinks is subject to scientific scrutiny (Moor, 1976, pp. 252–253). Jack Copeland summarizes some of the behaviorist accounts of Turing, and claims that "twenty-five years [after Moor's 1976 article], the lesson has still not been learned that there is no *definition* to be found in Turing's paper of 1950" (Copeland, 2001, p. 522). Daniel Dennett argues for this view in a bit more detail (Dennett, 1985, p. 6). Dennett does not provide textual evidence that Turing has this understand of his own test, but instead philosophical argument to convince the reader that this is the most reasonable understanding of the test. In particular, Dennett considers conditions under which the 'quick-probe assumption' (that success on the Turing test implies success on an indefinite number of other tasks) is false, and concludes that these conditions involve illicit constraints on the Turing test.

Later I will argue that Dennett's understanding of Turing's test can be given additional historical support. For now, I will turn to yet another set of reasons, closely related to the third set, to think that Turing ought not to have been considered a behaviorist, in any sense.

2.5. The third response reconsidered: the epistemic-limitation condition'

Elsewhere I have argued that Turing reveals, in his response to 'Lady Lovelace's objection,' (defined below) a commitment to a *necessary* condition for thought (Abramson, 2008). I call this the epistemic-limitation condition, and find evidence for it both in his 1950 paper, and in writings of Turing's unpublished during his lifetime.

In short, the epistemic-limitation condition states that for a computer to think, its behavior must be *unpredictable*, even by someone who has access to its programming. Methods of constructing machines to pass the test, by preprogramming in responses to specific questions, would cause failure of this necessary condition. This condition is mentioned by Turing in a number of places, most often in response to some form of Lady Lovelace's objection. Lady Lovelace's objection says that machines cannot think, since any behavior they display is the result of their programmer's intention for them to display that behavior. First I will provide a few of the texts in which Turing expresses this condition, and then make a few comments on this significance of the condition for Turing.

> ...Let us return for a moment to Lady Lovelace's objection, which stated that the machine can only do what we tell it to do...An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil's behavior. This should apply most strongly to the later education of a machine arising from a child-machine of well-tried design (or programme). This is in clear contrast with a normal procedure when using a machine to do computations: one's object is then to have a clear mental picture of the state of the machine at each moment in the computation. This object can only be achieved with a struggle. The view that 'the machine can only do what we know how to order it to do', appears strange in the face of this (Turing, 1950, pp. 454, 458–459).[9]
>
> It would be quite easy to arrange the experiences in such a way that they automatically caused the structure of the machine to build up into a previously intended form, and this would obviously be a gross form of cheating, almost on a par with having a man inside the machine (Turing, 1951b, p. 473).
>
> If we give the machine a programme which results in its doing something interesting which we had not anticipated, I should be inclined to say that the machine *had* originated something, rather than to claim that its behaviour was implicit in the programme, and therefore that the originality lies entirely with us (Turing, 1951a, p. 485).
>
> [As] soon as one can see the cause and effect working themselves out in the brain, one regards it as not being thinking, but a sort of unimaginative donkey-work. From this point of view one might be tempted to define thinking as consisting of 'those mental processes that we don't understand'. If this is right, then to make a thinking machine is to make one which does interesting things without our really understanding quite how it is done (Turing, Braithwaite, Jefferson, & Newman, 1952, p. 500).

The first and second of these quotations suggest at least two different ways that computers can be unpredictable.[10] Perhaps Turing intends merely that unpredictability of machines be faced by someone who has no knowledge of the program the machine is running. Another possibility, which I claim is supported by the other quotations, is that the computer runs a program that is unpredictable *even if one has access to the program itself*.

Notice that the first interpretation is quite weak. Suppose a programmer devises a clever algorithm for producing symphonies, each of which she wrote in a previous career as a composer. Then, so long as she doesn't tell anyone what the algorithm is (suppose that the programmer/composer takes their compositions and algorithm to the grave the moment the computer is switched on), the first interpretation suggests that Turing would be satisfied that this computer meets Lady Lovelace's objection. This is absurd; no one would, in this case, agree that the computer had originated the symphonies.

The first quotation does not hold up well independently under this interpretation. Turing points out that the knowledge in question of the computer, in the normal case, 'can only be achieved with a struggle.' If the 'mental picture' refers to a computational state, the programmer is in an excellent position to know this, either by producing a 'system dump' (a description of the total internal state of the computer) or working through the program and its input by hand. On the other hand, as programmers know very well, if 'mental picture' refers to a more general description of the gross functional properties of the program, a system dump will often be insufficient for such clarity. This is why debugging is such a 'struggle.'

The third quotation has as its goal, as do the previous two, to account for how machines can originate their own behavior, as opposed to merely acting as stand-ins for the ingenuity of the programmer. In this case, though, Turing explicitly supposes that *we give the machine a program*. One might wonder how someone can have a computer program in their hands that, when run, results in unanticipated behavior. The short answer is that, as Turing proved in his 1936 paper, under a reasonable assumption (the Church-Turing thesis), there will always be computers that are unpredictable even for someone who knows how they work.

The fourth quotation brings this point home. Rather than imagining that there is some lack of knowledge that makes brains and computers unpredictable, Turing imagines cases in which we peer

[9] Despite Turing's beginning his section on learning machines with an expressed interest in revisiting Lady Lovelace's objection, some seize upon isolated comments to conclude that for Turing, learning machines are merely an expedient path to building thinking machines. See, for example Davidson, 1990, p. 86).

[10] I am grateful to an anonymous referee for pointing out possible interpretations of these quotations and provoking clarification of their significance.

Please cite this article in press as: Abramson, D. Descartes' influence on Turing. *Studies in History and Philosophy of Science* (2011), doi:10.1016/j.shpsa.2011.09.004

inside ~~each~~ and lack *understanding* of what we see. Again, Turing is an expert on the existence of such cases. It would be very strange to hold the first interpretation of these quotations in attempting to explain the use of the word 'understanding' in the fourth quote. After all, those who observe the computer described above presenting its symphonies don't merely lack understanding of how the computer composes—they lack knowledge of how the computer operates altogether.

The epistemic-limitation condition names the lack of understanding that one has *even after seeing how a machine works*, that is, by observing its program. Learning computers provide a possible route to constructing such machines. In some cases, construction of a learning computer will fail to result in a machine that satisfies the epistemic-limitation condition. However, only building in previously understood programs in machines is *guaranteed* to result in machines that fail the epistemic-limitation condition. Thus the first two quotations emphasize the importance of not constructing machines that contain previously understood forms.

So, there is ample textual evidence that, in addition to providing a sufficient condition on intelligence, namely, passing the Turing test, Turing also holds a necessary condition on intelligence: the epistemic-limitation condition, as I have called it. An obvious question is, how can Turing consistently hold both of these? Doesn't calling the Turing test a sufficient condition *mean* that no other necessary conditions must hold for something that passes it?

In short, Turing is committed to the *empirical* claim that satisfaction of his sufficient condition (passing the Turing test) implies satisfaction of his necessary condition (the epistemic-limitation one) for having intelligence. To use a term from a widely cited and anthologized paper on the Turing test, Turing has what Ned Block calls a 'psychologistic' condition on thinking, but thinks that this condition will be satisfied by anything that passes the sufficient condition.[11]

The last response to Turing's claimed behaviorism is intimately connected to Dennett's response. In fact, Dennett's response can be thought of as the claim that the implication from satisfaction of the sufficient condition to satisfaction of the necessary condition can be justified on *a priori* grounds.[12] I won't offer an argument in support of that here. In the absence of such an argument on Turing's part, the parsimonious reading is that he simply believed a connection between his necessary and sufficient conditions for intelligence is likely, and worth testing.

2.6. Summary

Of the four responses to the claim that Turing offers a behavioral analysis of the possession of mental states, the last is the strongest. It does attribute to Turing the claim that passing the Turing test is a sufficient condition for intelligence, an apparently behavioral criterion. However, on the strength of considerable textual evidence, Turing believes that satisfaction of this behavioral criterion implies satisfaction of a non-behavioral criterion. Furthermore, Turing believes that this non-behavioral condition *must* be satisfied—is necessary for—having a mind.

So far I have been merely setting up the problem. Now that I have shown that Turing wasn't a behaviorist in any sense, one can ask: what influences was Turing acting under, if not his zeitgeist? I will show in the next section of the paper a significant influence for Turing in the formulation of his sufficient condition for intelligence.

3. A Possible source for the Turing test

3.1. Introduction

In arguing for Turing's commitment to his necessary condition I have suggested that the usual understanding of the Turing test is mistaken: it is not an expression of behaviorism. It bears explaining, then, what historical influences (if any) contributed to Turing's formulation of his test, and whether these provide additional insights into how to understand Turing's conception of his test.

Now I will show that his *sufficient* condition was not original to Turing, but taken, with light modification, from a significant figure in the history of philosophy. So, in the remainder of this paper, I will discuss the origin of the Turing test. First I review some hypotheses concerning what, if any, influences contributed to Turing's development of his test. Then I offer and justify a particular hypothesis. Finally, I try to show that the test, and its source, share deep commonalities. In short, I claim that Turing's test, and Descartes' so-called language test, are epistemologically analogous—they play similar roles for each in collecting information about whether some object thinks. I both appeal to existing interpretations of Descartes and Turing, and offer new historical evidence in support of this interpretation.

3.2. Descartes and Turing

Some commentators have tried to deduce the origin of the Turing test from an analysis of Turing's work. Here is part of an attempt by Hodges, in his biography of Turing:

> The discrete state machine, communicating by teleprinter alone, was like an ideal for [Turing's] own life, in which he would be left alone in a room of his own, to deal with the outside world solely by rational argument. It was the embodiment of a perfect J.S. Mill liberal, concentrating upon the free will and free speech of the individual. From this point of view, his model was a natural development of the argument for his definition of 'computable' that he had framed in 1936, the one in which the Turing machine was to emulate anything done by the individual mind, working on pieces of paper. (Hodges, 1983, p. 425)

So, the Turing test, according to Hodges, is the confluence of Turing's views on the equivalence of effectively computable functions and Turing computable functions, and his own personal political and social temperament. In a similar vein, A.K. Dewdney writes '[Turing's] proposal [for the Test] was the essence of British fair play: A human judge would interact with either a computer or a human and then guess which was which' (Dewdney, 1992, p. 30).

Daniel Dennett is, to my knowledge, the only person to have even considered the possibility that Turing may have been inspired by previous philosophical thinking on the difference between minds and machines. In the same article in which he denies that Turing is a behaviorist, Dennett writes, 'Perhaps [Turing] was inspired by Descartes, who in his *Discourse on Method*, plausibly argued that there was no more demanding test of human mentality than the capacity to hold an intelligent conversation' (Dennett, 1985, pp. 5–6). In the relevant passage, Descartes argues that there are sure ways to distinguish beings that think from mere machines. I will quote a slightly longer passage than Dennett does from the *Discourse*.

[11] See Block (1981). Clearly, Ned Block does not find this view in Turing's own work. However, Block's paper can be understood as a defense of the epistemic-limitation condition, together with an argument that its relationship to Turing's sufficient condition must be contingent, not necessary.

[12] Recently, Stuart Shieber has offered a sustained argument that the connection between satisfaction of plausible versions of Block's psychologistic requirement and passing the Turing test can be justified on mildly *empirical* grounds. See Shieber (2007, p. 709).

Please cite this article in press as: Abramson, D. Descartes' influence on Turing. *Studies in History and Philosophy of Science* (2011), doi:10.1016/j.shpsa.2011.09.004

> . . . if any such machines had the organs and outward shape of a monkey or of some other animal that lacks reason, we should have no means of knowing that they did not possess entirely the same nature as these animals; whereas if any such machines bore a resemblance to our bodies and imitated our actions as closely as possible for all practical purposes, we should still have two very certain means of recognizing that they were not real men. The first is that they could never use words, or put together other signs, as we do in order to declare our thoughts to others. For 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). 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. Secondly, 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 (Descartes, 1637, pp. 139–140).

First, Descartes seems to think that, for machines lacking rationality, identical stimuli must give rise to identical responses ('if you touch it one spot, it asks what you want . . . But it is not conceivable that such a machine should produce different arrangements . . .'). Second, Descartes seems to think that once a machine has been assembled, there is a fixed, finite number of circumstances it can behave appropriately in ('these organs need some particular disposition for each particular action . . .'). The second test is like the first, but involves observing an open ended variety of abilities to accomplish physical, as opposed to verbal tasks.

The two related limitations of machines just mentioned preclude, on Descartes' account, the ability of a machine to acquire new dispositions, either for improving responses to circumstances it is ill-suited to in its beginning, or for circumstances it is initially unable to respond to at all. Descartes' reasoning thus leads naturally to Lady Lovelace's objection.

3.3. Descartes, Turing, and irony

Many of us, in trying to motivate the idea of the Turing test to students or colleagues, present these comments from the *Discourse* as a tonic for the complaint that Turing was an unreflective behaviorist. In fact, Jack Copeland, who also identifies precursors of the Turing test in the writings of Descartes and the Cartesian de Cordemoy, writes "The idea that the ability to use language is the hallmark of a thinking being has a long history. Ironically, the seventeenth century French philosopher Rene Descartes proposed conversation as a sure way of distinguishing any machine, no matter how subtle, from a genuine thinking being" (Copeland, 1993, pp. 38–39). The irony, I take it, is that Turing, an apparent materialist about mind, and Descartes, a dualist, agree on how we can determine that machines do or don't have minds.

However, other commentators have suggested, alternatively, that Descartes and Turing have distinct motivations for offering their criteria for the presence of mind, and that their tests are not even comparable (for example, Chomsky (2004)). First I will establish a likely influence, for Turing, in formulating his test. Then I will examine each of these claims concerning the similarities between Descartes and Turing.

3.4. The Turing test: an adapted language test

In this section I will argue that Turing's primary source of inspiration for the Turing test was not his British upbringing, social idiosyncrasies, nor even his views in computability theory. Rather, Turing's test finds its likely origin in, yes, Descartes' comments in the *Discourse*.

It is widely known that Turing, in writing his 1950 paper, read and responded to a paper called 'The Mind of Mechanical Man' by the neurosurgeon Geoffrey Jefferson. This paper was delivered as the Lister Oration at the Royal College of Surgeons of England on June 9, 1949. In responding to what he calls 'The Argument from Consciousness' against the possibility of machine thought, Turing quotes Jefferson at length. The online Digital Turing Archive contains an image of the page from the preprint of Jefferson's paper, in Turing's possession as he was writing his 1950 *Mind* paper, that is the source of this quote. In the margin, next to the passage from Jefferson that Turing quotes, there is a heavy line made in colored pencil. (http://www.turingarchive.org/viewer/?id=504&title=a)

The King's College Archive, at Cambridge University, in notes recorded when this preprint was donated to it, indicates that annotations to the preprint were in Turing's hand. The Archive's catalog entry describes the preprint in a batch of documents, left to the Archive by Robin Gandy, as having "annotations by AMT (Alan Turing)." (http://www.turingarchive.org/browse.php/B/33-57)

However, in examining Turing's preprint of the Jefferson paper in the physical archive, I found a second heavy line—so heavy, that the indentation from the pencil carries through 5 pages. Here is the *other* passage that Turing annotated:

> Descartes made the point, and a basic one it is, that a parrot repeated only what it had been taught and only a fragment of that; it never uses words to express its own thoughts. If, he goes on to say, on the one hand one had a machine that had the shape and appearance of a monkey or other animal without a reasoning soul (i.e., without a human mind) there would be no means of knowing which was the counterfeit. On the other hand, if there was a machine that appeared to be a man, and imitated his actions so far as it would be possible to do so, we should always have two very certain means of recognizing the deceit. First, the machine could not use words as we do to declare our thoughts to others. Secondly, although like some animals they might show more industry than we do, and do some things better than we, yet they would act without knowledge of what they were about simply by the arrangement of their organs, their mechanisms, each particularly designed for each particular action (cp. Karel Čapek's *Robots*). Descartes concluded: 'From which it comes that it is morally impossible that there be enough diversity in a machine for it to be able to act in all the occurrences of life in the same way that our reason would cause us to act. By these means we can recognize the difference between man and beasts.' He could even conceive a machine that might speak and, if touched in one spot, might ask what one wanted—if touched in another that it would cry out that it hurt, and similar things. But he could not conceive of an automaton of sufficient diversity to respond to the sense of all that could be said in its presence. It would fail because it had no mind (Jefferson, 1949, p. 1106).

It is therefore extremely likely that Turing was aware of Descartes' views on the claimed in-principle difference between minds and machines. Descartes' views at least helped crystallize Turing's own conception of the Turing test, and at most presented him with

Please cite this article in press as: Abramson, D. Descartes' influence on Turing. *Studies in History and Philosophy of Science* (2011), doi:10.1016/j.shpsa.2011.09.004

the idea *in toto*. The King's College Archive contains one preprint of the paper that Turing read and quoted from. That preprint contains two annotations that, according to the Archive, are in Turing's own hand. One annotation is of a passage explicitly quoted by Turing; the other is an annotation of an expression of the central idea of the 1950 paper: that thinking things can be distinguished from non-thinking things by a flexible ability to use natural language.

Jefferson's paper not only paraphrases Descartes' views, but endorses them. Jefferson asserts a materialist view, and presents his paper as an attempt to reject Descartes' dualism concerning brain and mental function. He states that the notion that minds are physical objects seems to offend both our sense of the richness of mental states, and our ethical and political self image. However, Jefferson goes on to try to show that although minds are physical things, no computer could ever pass Descartes' language test. Jefferson's reason for thinking this is that he is "quite sure that the extreme variety, flexibility, and complexity of nervous mechanisms are greatly underestimated by the physicists, who naturally omit everything unfavourable to a point of view" (Jefferson, 1949, p. 1110).

In consideration of 'the argument from consciousness,' Turing quotes Jefferson to the effect that machines cannot think because, no matter what they do, they will lack accompanying emotions and feeling. In the quoted passage, Jefferson lapses into a position inconsistent with the one that appears elsewhere in his paper. That is, Jefferson claims that even if a computer could perform language tasks, one could still question whether or not consciousness or reason were behind the expressions. This comment is made despite Jefferson's approving presentation of Descartes elsewhere in the paper. Turing's selective presentation of Jefferson's views may have prevented later readers of Turing from investigating Descartes' influence on Turing, via Jefferson, further.

Jefferson, to use contemporary terms from cognitive science, rejects multiple realizability and adopts something like the dynamical hypothesis, claiming that machines can only be imperfect mimics of the brain: "however [the human brain's] functions may be mimicked my machines, it remains itself and is unique in Nature" (Jefferson, 1949, p. 1106).

In the next section I will argue that Descartes and Turing both understand their own tests in the same way: as empirical hypotheses concerning a theoretical commitment to the nature of mind. That is, I will show that each of them thinks that satisfaction of their test implies the presence of some inner, necessary condition for a mind. But, we will see, their commitment to this implication is subject to empirical investigation.

3.5. Moral Impossibility

First I will present a widely held understanding of Descartes' language test, argue briefly in favor of it against an alternative, and then show that this yields a deep epistemic commonality between the two tests. To begin, let us pose the difficult historical question: if Descartes was made aware both of the Turing test, and a machine that passed it, would he be compelled to abandon his view that mental substances and physical substances are distinct? Or, at least, we can pose the slightly less difficult question: what view would be consistent with Descartes' remarks on the language test?

Descartes' comments in the *Discourse* use qualifications to describe the possibility of machines that pass the language test, but do not possess reason (and therefore a soul). Descartes describes the most complicated machines *we can conceive of*, and then says that they 'could never' perform as even the stupidest humans do with natural language (Descartes, 1637, p. 140). Finally, Descartes says that for any machine that is as close to a person 'as possible for practical purposes', we would have 'very certain' ways of telling it apart from real people.

I want to suppose, then, that the qualification of 'moral impossibility' applies to the case of a machine that passes the language test. Now, it is unlikely that Turing would have been aware of what this term meant for Descartes. So we cannot argue from Turing's exposure to Descartes' view that he understood the Turing test to similarly provide the same level of certainty. On the other hand, our question is served is by examining the definition of this technical phrase.

In *Principles of Philosophy*, Part Four, Descartes writes

> It would be disingenuous, however, not to point out that some things are considered as morally certain, that is, as having sufficient certainty for application to ordinary life, even though they may be uncertain in relation to the absolute power of God. <Thus those who have never been in Rome have no doubt that it is a town in Italy, even though it could be the case that everyone who has told them this has been deceiving them> (Descartes, 1644, pp. 289–90).

'Morally certain' in this passage means 'not absolutely certain'. It is therefore reasonable to interpret 'morally impossible' in the passage from the *Discourse* as 'not absolutely impossible'. Then there are at least two different interpretations of Descartes' remarks in the *Discourse*, one of which allows him to maintain his position even after being presented with the Turing test and a machine that passes it, and another that does not.

I call the first interpretation the 'conditional probability' one. According to it, the probability that a given object has a soul, given the evidence that it passes the Turing test, is extremely high. However, this evidence can be defeated on the discovery that the object is a mere machine. Let us interpret the modal operator $\Box$ epistemically. That is, for any sentence P, $\Box P$ will mean 'I believe it to be nearly impossible (but not absolutely impossible) that P is false.' In formal terms, the first reading of Descartes' commitment to the language test/Turing test can be expressed as

$$\forall x \Box(PassesTheTuringTest(x) \rightarrow NotAMachine(x))$$

I will call the second interpretation the 'Turing' interpretation. According to the Turing interpretation, Descartes holds a universal claim with near certainty. The claim is

$$\Box \forall x(PassesTheTuringTest(x) \rightarrow NotAMachine(x))$$

Notice that on this latter interpretation, Descartes is committed to the material conditional that passing the Turing test implies the presence of reason, as opposed to mere mechanism. However, although his level of commitment is high, it is possible that Descartes could be mistaken, in which case there is *no* implication from the ability to use natural language to the presence of some non-mechanical process.

The Turing interpretation is supported by some readers of Descartes. For example, in his analysis of Cartesian dualism, John Cottingham claims that Descartes provides divergent arguments that are in tension with one another. In particular, Cottingham claims that the language test displays a 'scientific' motivation for dualism which is defeasible on possible evidence, whereas Descartes' metaphysical argument (involving an argument for the separability of body and mind) is not subject to empirical evidence (Cottingham, 1992).

Cottingham neatly ties together Lady Lovelace's objection and Descartes' views, by suggesting that Descartes merely assumes that no machine can be built that is unpredictable by its creator (Cottingham, 1992, p. 250). Presumably, the creator of a mechanical device can see first hand the assemblage of organs with determined dispositions that will produce the device's behaviors. If a machine passes the language test, then it will have to perform in ways that its creator cannot anticipate, since otherwise the programmer will have to imagine all of the indefinite things the machine can do.

Please cite this article in press as: Abramson, D. Descartes' influence on Turing. *Studies in History and Philosophy of Science* (2011), doi:10.1016/j.shpsa.2011.09.004

On Cottingham's view, Descartes believes that the language test is sufficient for distinguishing thinking beings from machines *precisely because Descartes cannot imagine that a machine will ever satisfy the epistemic-limitation condition on intelligence*.

Note that the Turing interpretation of Descartes' views of his language test leaves open what Descartes would actually do if he were confronted by a talking machine. Cottingham's interpretation implies that Descartes would give up his test in the face of an apparent counterexample. On the other hand, Keith Gunderson writes "Even if 'another Prometheus' made a highly convincing talking mechanical man, I believe it is more likely that Descartes would rather have claimed that a generous God had granted the clever fellow an extra soul to go with his invention, than submit to the conclusion that we had no soul at all" (Gunderson, 1971, p. 34). Whether or not Gunderson is right about what Descartes would do, Gunderson in this passage attributes the Turing interpretation to Descartes. That is, Gunderson thinks that for Descartes, the alternative hypothesis, that the language test is insufficient, is less probable than some alternative involving a soul. Gunderson does not think that Descartes has open to him the possibility that he has been confronted by the rare machine that can converse without having a soul.

Cottingham most clearly endorses the Turing interpretation of Descartes' language test over the conditional probability interpretation with his claim that Descartes believes strongly that the limits of physics would prevent any object, operating according to purely mechanical principles, from having the ability to converse in natural language (Cottingham, 1992, p. 252).

I will now provide some brief additional philosophical considerations in support of the Turing interpretation of the language test over the conditional probability interpretation.

Suppose that a machine can be built that passes the Turing test, and that Descartes is presented with it. Given a manufacturing technique that can produce a single machine that passes a Turing test, many more such machines can be created by just copying the first one. So, if the very small likelihood obtains that a machine exists that passes the Turing test, one can, conceivably, revise the probability of some object's being a non-thinking possessor of natural language ability to approach any measure of likelihood. Such scenarios include ones in which the machines being manufactured take control of the manufacturing process. Descartes clearly does not intend his commitment to the moral impossibility of language-using machines to rely on individual empirical observations; rather, admitting the existence of a language-using machine, even in a single instance, requires rejecting whole networks of beliefs and commitments. Therefore, even given the qualifications that Descartes offers, he would be compelled to at least revisit his dualism if Turing's assertion, that a single machine can be built that passes the Turing test, is correct.

Here is an analogous commitment for Turing that both makes clear that he is not a behaviorist, and highlights the similarity between his understanding of his test and Descartes' (*SatisfiesEpLim*(x) means x satisfies the epistemic-limitation condition):

$$\Box\forall x(PassesTheTuringTest(x) \rightarrow SatisfiesEpLim(x))$$

This is the view that I interpret Turing as holding. He holds, with a high degree of certainty, that satisfaction of his sufficient condition for intelligence implies satisfaction of his necessary condition.

Consider now a conditional probability position on the relationship between the necessary and sufficient conditions for intelligence that Turing offers:

$$\forall x\Box(PassesTheTuringTest(x) \rightarrow SatisfiesEpLim(x))$$

Turing cannot hold the weaker, conditional probability view, and maintain a necessary condition on intelligence, while holding that passing the Turing test is a sufficient condition for having a mind. Given the evidence for Turing's commitment to his necessary condition on intelligence, I claim that the first interpretation of Turing's position, analogous to what I have called the 'Turing' interpretation of Descartes, is more plausible. In both cases, there is a strong commitment to a relationship between possession of properties that may be falsified through further empirical and theoretical investigation. Both Turing and Descartes hold their test in the same status, *mutatis mutandis* for each's necessary condition for having a mind.

There is a limit, of course, to how similar Descartes and Turing can be in their understanding of their tests. Descartes subscribes to a necessary, internal *and* sufficient condition for the possession of intelligence: the having of an immaterial soul. Descartes believes, though, that we can't observe souls directly in others, and must rely on the test to detect their presence. Turing, on the other hand, has his test (constituting a scientific commitment rather than a statement of a behaviorist condition for intelligence), but no independent sufficient condition for mind. Perhaps we can make sense, then, of Turing's distaste for discussions of the meaning of terms like 'thinking,' and his (merely apparent) suggestions that the imitation game operationalizes intelligence. By rejecting dualism, Turing has no alternative, internal, sufficient condition for intelligence. But, Turing encourages us, this gap in our understanding need not preclude scientific inquiry.

4. Conclusion

So, I believe there is ample evidence that Turing at least conceived of his own test as fulfilling just the purpose that Descartes' fulfilled for him. I have argued for this by presenting extant interpretations of Descartes, analysis of Turing's texts, and philosophical analysis of the views of each.

Turing is in a dialogue spanning centuries in which he is presented with the view that, due to some hidden property, humans are able to engage in natural language conversations, but computers aren't. Faith that a machine can be built that passes the Turing test constitutes a denial of this claim, together with the belief that such a machine can be built lacking any special physical or metaphysical property. Viewed this way, the Turing test is not a merely rhetorical tool designed to influence scientific or social commitments, but instead a concrete method for settling philosophical disputes over what can be taken to indicate the presence of a mind. I claim that the Turing test and Descartes' language test fulfill exactly the same purpose—testing for the presence of some property that is necessary for mind, and claimed by some to be unimplementable in mere machines.

Turing was aware of Descartes' language test, and likely was inspired by this to come up with the Turing test. Finally, on a defensible reading of both Descartes and Turing, performance in natural language contexts indicates to both a hidden, necessary property for intelligence.

Acknowledgements

This research was supported by a Dalhousie Faculty of Arts and Social Sciences Research Development Fund grant, from funds provided by the Social Sciences and Humanities Research Council of Canada. I am grateful to helpful feedback from an anonymous referee, participants at the 2008 meeting of the Canadian Society for the History of Philosophy and Science, and participants of the Dalhousie Philosophy Colloquium Series. I am also grateful to Joy Abramson, Duncan MacIntosh, Tom Vinci, Sara Parks and Lisa Kretz for encouragement and discussion while working through previous versions of this paper.

Please cite this article in press as: Abramson, D. Descartes' influence on Turing. *Studies in History and Philosophy of Science* (2011), doi:10.1016/j.shpsa.2011.09.004

References

Anderson, A. R. (Ed.). (1964). *Minds and machines*. Prentice-Hall.
Abramson, D. (2008). Turing's responses to two objections. *Mind and Machines, 18*(2), 147–167.
Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review, 90*(1), 5–43.
Chomsky, N. (2004). Turing on the "imitation game". In S. Shieber (Ed.), *The Turing test: Verbal behavior as the hallmark of intelligence* (pp. 317–321). MIT Press.
Copeland, J. (1993). *Artificial intelligence: A philosophical introduction*. Blackwell Publishers.
Copeland, B. J. (2001). The Turing test. *Minds and Machines, 10*, 519–539.
Cottingham, J. (1992). Cartesian dualism: Theology, metaphysics, and science. In J. Cottingham (Ed.), *The Cambridge companion to descartes*. Cambridge University Press.
Davidson, D. (2004/1990). Turing's test. In *Problems of rationality*. Oxford University Press.
Dennett, D. (1998/1985). Can machines think? With postscripts 1985 and 1997. In *Brainchildren: Essays on designing minds*. MIT Press.
Descartes, R. (1985/1637). Discourse on method. In *The Philosophical Writings of Descartes* (Vol. I)(John Cottingham, Robert Stoothoff, Dugald Murdoch, Trans.). Cambridge University Press.
Descartes, R. (1985/1644). Principles of philosophy. In *The Philosophical Writings of Descartes* (Vol. I)(John Cottingham, Robert Stoothoff, Dugald Murdoch, Trans.). Cambridge University Press.
Dewdney, A. K. (1992). Turing test. *Scientific American, 266*(1), 30–31.
Gunderson, K. (1971). *Mentality and machines*. Doubleday.
Hodges, A. (1983). *Alan Turing: The enigma*. Burnett Books.
Jefferson, G. (1949). The mind of mechanical man. *British Medical Journal, 1*(4616), 1105–1110.
Leiber, J. (1995). On *Turing's* Turing test and why the *matter* matters. *Synthese, 104*, 59–69.
Moor, J. H. (1976). An analysis of the Turing test. *Philosophical Studies, 30*, 249–257.
Piccinini, G. (2000). Turing's rules for the imitation game. *Minds and Machines, 10*(4), 573–582.
Scriven, M. (1964). The mechanical concept of mind (with postscript). In A. R. Anderson (Ed.), *Minds and machines* (pp. 31–42). Prentice-Hall.
Searle, J. (2004). *Mind*. Oxford University Press.
Shieber, S. (2007). The Turing test as interactive proof. *Noûs, 41*(4), 686–713.
Sterrett, S. (2000). Turing's two tests for intelligence. *Minds and Machines, 10*(4), 541–559.
Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.
Turing, A. (2004a). Can digital computers think? In B. J. Copeland (Ed.), *The essential Turing*. Oxford University Press.
Turing, A. (2004b). Intelligent machinery, a heretical theory. In B. J. Copeland (Ed.), *The essential Turing*. Oxford University Press.
Turing, A., Braithwaite, L. C., Jefferson, A. A., & Newman, E. (2004). Can an automatic calculating machine be said to think? In B. J. Copeland (Ed.), *The essential Turing*. Oxford University Press.

Please cite this article in press as: Abramson, D. Descartes' influence on Turing. *Studies in History and Philosophy of Science* (2011), doi:10.1016/j.shpsa.2011.09.004