

Man, Machine, and Interpretation.

Donald Davidson on Turing's Test

Stefan Rieglernik

Stefan.Rieglernik@wu.ac.at

In the year 1950 Alan Turing published his influential paper “Computing Machinery and Intelligence”, where he presented at first the idea of a machine, which is able “*to carry out any operations which could be done by a human [being]*” and insofar “*can in fact mimic the actions*” of a human being “*very closely*” (Turing 1950). According to Donald Davidson the remarkable feature of this idea is that it presupposes what a lot of people in the following discussion have missed as a decisive aspect of the behaviour of the machine: namely to understand what it does, i.e. to have an understanding of the meaning of the signs which the machine manipulates. This assessment is in some way similar to a lot of doubts about the fruitfulness of Turing's test by philosophers concerned with language and mind. In contrast to John Searle, as one of the prominent critics of Turing's idea, for Davidson this does not show the deficiency of the idea of the machine as a criterion for having thoughts. It rather proves that questions concerning the conditions of having thoughts depend on questions concerning the conditions of attributing thoughts. So when Donald Davidson explicitly focused on Turing's Test (1990) as a thought experiment his assessment was not at all completely negative, but rather ambiguous: on the one hand he regards the test as inadequate, on the other hand he concedes that it reveals some important aspects concerning the attribution of thought. My contribution to the still ongoing discussion about Turing's Test is meant to emphasize the significance of the argumentation by Donald Davidson as a philosopher who has been regrettably neglected in the assessment of Turing's test.

*

The philosophical relevance of Turing's test does not seem apparent at first sight, since one could argue that nothing follows from asking people, i.e. asking whether they believe that some thing thinks: if on the contrary one believes it does, then the fact that an object “succeeds” in performing in a way, that is assessed by the interrogator as a way human beings act and talk, thereby scoring a certain amount of points, then one could rightly draw the conclusion that the object is what other people take it to be meaning in this context that the object *thinks*, because human beings think that it thinks. This is especially plausible since “alternatives”, like a “direct” way to another person's thought is not available. Thus one could point out that Turing's test serves indeed as a proof that a machine can mimic the behaviour of a human being, showing thereby that Turing's test is an adequate substitution for his initial question “Can machines think?”.

The focus on this “third person point of view” is one reason for which Turing’s test has been criticized. Davidson stresses the question of the presupposed truth of behaviourism and points out, that since “*the Test is to see whether, and on what basis, normal human judges are able or willing to assign mental attributes to machines*”, behaviourism “*in the broadest sense must be assumed*”. According to Davidson, this cannot be objected to Turing’s Test, since “*the same question will apply equally to everyone (else)*” (Davidson 1990): no one has access to the thoughts of a thinking being without observing the behaviour. Talking about thought requires adopting a third person point of view. Davidson is strictly against being able to settle the question whether an object can think *without* being able to say anything about the *content* of her/his thoughts. I think this is important to mention here because in the history of philosophy many attempts have been made to explain the nature of thought that ignore the fact that thoughts without content are “empty”. A view opposed to this ends always up with comparing thoughts with rivers, streams (anything in a flux), or with a theatre. What “unifies” these metaphors is the fact that essential features of thoughts cannot be explained, or even worse, as Mras shows in her work on Nietzsche and Wittgenstein, these metaphors are rather misleading than helpful. The possibility of ignoring the content is excluded in the scenario of Turing’s test – “*any evidence that thinking is going on will have to be evidence that particular thoughts are present*” (Davidson 1990). Davidson deems this as a particularity of Turing’s test – he contrasts this with normal circumstances where it is not necessary to discover *what* a human thinks to know *that* it thinks “*by being told, or knowing in some other way that it is a man or woman*” (Davidson 1990). Although Davidson weakens this by holding that these “*are fallible ways of knowing*”, this line of argument makes Davidson’s position vulnerable to arguments that Turing’s test is a test for *human* intelligence only.

*

In order to show that even from a third person point of view a sceptical attitude towards the equation of human and artificial intelligence is justified, I am going to examine the conditions of ascribing meaning to purely mechanically produced signs. Considering the fact, that Turing’s test is an artificial situation, Davidson admits that many aspects of human behaviour have to be removed from consideration leaving solely the occurrence of signs as expressions of what humans do. In other words, the occurrence of the signs alone could enable the interrogator to decide whether the sign-producing object has thoughts or not, which comes down that “*[t]he Test makes meaningful verbal responses the essential mark of thought*” (Davidson 1990).

The focus of Davidson’s argument lies at first on the interdependency of “having thoughts” and “the signs” or better, “the meaning of the signs”. He writes, “*if a response has any linguistic meaning at all, there must be a thoughtful cause. But the computer is*

the thoughtful cause of a response only if it means something by the words it produces” (Davidson 1990). Only if the interrogator can identify the signs appearing on the screen as linguistic utterances *and* if he is able to find out what they mean, he knows the content of the thoughts expressed and consequently, *that* the machine thinks. This line of reasoning reminds one of Searle’s thought experiment known as the “Chinese Room Argument”. For Searle the fact that signs occur on the screen does not show that the machine that “does this process” exhibits thereby a meaningful behaviour. What computers according to Searle do is at most a formal manipulation of signs. Davidson argues in a similar way: “[...] Turing has his computer produce what seem to be English sentences. But of course the computer might not be speaking English. Those sentences might have entirely different meanings for the computer than they have in English; or they might have no meaning at all” (Davidson 1990).

The question of whether there is a relation between thought and talk and whether the relationship is necessary are here of minor interest; the pivotal question is now whether the interrogator can find out the meanings of the “signs” produced by the machine. This is essential, since the signs are the sole evidence to find out *what* the object thinks and consequently, *that* it thinks.

*

One might argue here that too much is taken for granted, since how can we know that the results of machine’s operations are to be understood as *signs* at all? The only way to settle this question is to take up the effort to understand the expressions of the machine, i.e., to find out the meaning if what might be taken as signs. At first it seems that there cannot be any difficulty involved in this task, since the interrogator could interpret the signs appearing on the screen just as every other utterance. A common objection to this line of reasoning is that a human being has programmed, e.g. an ATM, so the signs appearing on the screen are purely “formal symbol manipulations”: the signs are not produced by a thoughtful machine but by a programmer. Davidson writes: “[...] *it is the programmer who has supplied the semantics, who has understood English, and has given meaning to the words produced by the object*” (Davidson 1990). Nonetheless one could still ask: how can we exclude that it is the machine that caused the utterances occurring on the screen? The peculiarity of this question is that it cannot find an answer without presupposing the question already having been answered.

For this reason, Davidson modifies Turing’s Test. The new scenario has a single object under scrutiny. The interrogator now has to find out whether the object thinks or not by trying to interpret the signs produced. This test is different and so are the requirements in order to satisfy it, but bearing in mind the fact that Turing’s Test eliminates the possibility of telling whether an object thinks without knowing what it thinks, Davidson’s Test is as

suitable as Turing's, with a minor adjustment: since Davidson understands Turing's test as a way to explore the nature of thought, the modified test is more suitable "*if what we want to know is what our (the interrogator's) criteria for thought are*" (Davidson 1990). For reasons of alienation, it would be useful to have the machine uttered words in a language unknown to the interrogator, but as Davidson holds elsewhere, "[t]he problem of interpretation is domestic as well as foreign: it surfaces for speakers of the same language in the form of the question, how can it be determined that the language is the same?" (Davidson 1973).

If we assume that the signs on the screen are English sentences (Davidson cites a correct syntax and correct relations between questions and answers in favour of this assumption and argues that "*we have no better evidence if the object were an English-speaking person*"), the question is whether the interrogator is able to interpret these signs. Davidson denies this idea because the interrogator could not have a "*clue to the semantics of the object*". The interrogator cannot tell, in spite of the correct syntax and the harmonization of questions and answers, what the signs mean since "[t]here is no way he [the interrogator] can determine the connection between the words that appear on the object's screen and events and things in the world" (Davidson 1990). But why not? Davidson mentions two methods of finding out what people mean by what they say: the "indirect" method, whereby one identifies a person who learned a language "*through the usual conditional process, which connect things and events with words*" (Davidson 1990), and the "direct" method, which is the sole available in the test scenario, where "*we can discover these connections directly, through observing relevant causal interactions between the speaker, the world, and the speaker's audience*" (Davidson 1990). The latter is very much the same as Davidson's thought experiment "radical interpretation", which is intended to pinpoint the required knowledge for being able to interpret linguistic utterances. The radical interpreter can rely on no prior knowledge of the speaker's beliefs or of the meanings of his utterances: he has to interpret the utterances *from scratch*. The only material available to the interpreter is the speaker's behaviour towards objects and events. To describe the process of radical interpretation and, simultaneously, the attribution of thoughts, Davidson suggests to understand the attribution of thoughts in the model of a triangular structure: "*If the two people [...] note each other's reactions (in the case of language, verbal reactions), each can correlate these observed reactions with his or her stimuli from the world. [...] The triangle which gives content to thought and speech is complete.*" (Davidson 1991).

*

Davidson thus regards Turing's Test as inadequate not because "*the Test restricts the available evidence to what can be observed from outside*" but "*because it does not allow*

enough of what is outside to be observed” (Davidson 1990). The interrogator is not able to find out what the signs on the screen mean, since “[w]hat is needed is evidence that the object uses its words to refer to things in the world, that its predicates are true of things in the world, that it knows the truth conditions of its sentences” (Davidson 1990). In consequence, Davidson suggests a modification of Turing’s Test: “[t]he object must be brought into the open so that its causal connections with the rest of the world as well as with the interrogator can be observed by the interrogator” (Davidson 1990). However, it is an open question whether a suitable test scenario along these lines could be designed at all. So either the “object brought into the open” must resemble a person in all important aspects, or if it does not, it is clear from the start that the machine “described as mimicking the actions of a human being very closely” does so because it is meant to deceive the interrogator.

References

- Davidson, D. (1973): Radical Interpretation. In: *Dialectica*, 27, 212-328.
- Davidson, D. (1990): Turing’s Test. In: Newton-Smith, W.H. / Wilkes, K.V. (eds): *Modelling the Mind*. Oxford University Press. 1-11.
- Davidson, D. (1991): Three Varieties of Knowledge. In: Phillips Griffiths, A. (ed): *A.J. Ayer Memorial Essays: Royal Institute of Philosophy Supplement*. Cambridge University Press.
- Mras, G. (2010): Nietzsche and Wittgenstein. Taking on different perspectives. In: Ramharter, E. (ed): *Influences on Wittgenstein*. (forthcoming 2010).
- Searle, J.R.: *Intentionality. An essay in the philosophy of mind*. Cambridge: Cambridge University Press. 1983.
- Turing, A.M. (1950): Computing machinery and intelligence. In: *Mind*, 49, 433-460.