Wittgensteinian Perspectives on the Turing Test

Ondřej Beran

Department of Philosophy, University of West Bohemia

This paper discusses some difficulties in understanding the Turing test (TT). It emphasizes the importance of distinguishing between conceptual and empirical perspectives and highlights the former as introducing more serious problems for the TT. Some objections against the Turingian framework stemming from the later Wittgenstein's philosophy are exposed. The following serious problems are examined: 1) It considers a unique and exclusive criterion for thinking which amounts to their identification; 2) it misidentifies the relationship of speaking to thinking as that of a criterion; 3) it neglects the "natural" course of the development in semantics. However, these considerations suggest only that it is problematic to label a successful chatbot as a "thinking entity" without further qualifications, but not necessarily and once and for all incorrect. Philosophy has only little to say about the technical possibility of creating such an effective program.

Keywords: Turing test, artificial intelligence (AI), thinking, Wittgenstein

1. Introduction

In this paper, I will show the reasons for some philosophers' doubts about the Turing test (TT) and its naïve, literal understanding present in some endeavors at passing the TT. I will adopt here the analytical perspective on conceptual issues, based mainly on the later Wittgenstein, but not confined to his rather scarce and scattered remarks. In section 2, I will distinguish between empirical and conceptual readings of Alan Turing's test proposal and of some of its most important criticisms and I will recapitulate these briefly. Section 3 presents two objections (made by Carlo Penco and Stuart Shanker) using Wittgenstein's remarks rather closely concerning the implicit (incorrect) notion of language used in the general Turingian account. I will argue that though these objections make a good point, they do not necessarily undermine the whole framework. In section 3, I will use some more

Corresponding author's address: Ondřej Beran, Department of Philosophy, University of West Bohemia, Sedláčkova 19, Plzeň 306 14, Czech Republic. Email: jerdnonareb@seznam.cz.

Studia Philosophica Estonica (2014) **7.1**, 35–57
Published online: April 2014
Online ISSN: 1736–5899
www.spe.ut.ee
http://dx.doi.org/10.12697/spe.2014.7.1.02

general analytical tools provided by Wittgenstein, not only his direct (rather narrow and specific) criticisms on Turing, to show that problems arise from there being a unique and exclusive criterion for something and from confusing the criterion with the phenomenon it stands for (a problem enabled but bypassed by Turing himself). The overall ambition of the paper is not to estimate the prospects of the competitors in the TT—but to evaluate them from an empirical and technical point of view (that reaches beyond my expertise). Rather, I will suggest what conceptual implications the present attempts yield and what possible future success would mean. Finally I will suggest a few considerations that could help illuminate some conceptual shortcomings in understanding the achievements of chatbot programming.

2. Two Points of View on the Turing Test

In the more than 60 years since its publication, many interpretations of Alan Turing's (1950) idea of a test for AI, as well as many arguments both in its defense and—mainly—criticizing it have occurred (for a thorough and informative survey up to 2000, see Saygin et al. 2000). I will focus here on one possible reading dichotomy, interesting for philosophy. The Turing test (TT) is in itself a proposal of a method of answering (in fact, bypassing) the question, "Can machines think?" But since the idea of tracing the thinking machine in terms of a talking machine (chatbot) combines *conceptual* with *empirical* elements, the question addressed by the TT falls into two different questions.

1. The first question can be rephrased, e.g. as follows: "Does our term 'thinking' (or 'intelligence') also meaningfully admit machines as thinking entities?" This—conceptual—reading is characteristic of the Wittgensteinian criticisms of Turing such as Shanker (1998).

In order to answer it, we have to ask what 'thinking' means and consider that the terms of our languages are strongly interlinked to one another, and that applying one entails applying others also. We thus apply, on the face of it justly, the term 'speaking' to parrots, but we can hardly ascribe 'storytelling' to these creatures—because this term works within our language only in connection with 'to resume the plotline', 'to continue narrating where one has stopped' (continue the *story*, not the exact word sequence), etc. At first sight one is tempted to admit the idea of storytelling parrots because some talented individuals among them are capable of pronouncing whole sentences that sometimes even look as if one sentence takes up the preceding one meaningfully, etc. On second inspection a parrot can tell a story only if it makes sense to ask and expect the bird to summarize the actual plotline, continue the story after a while, or re-tell a specified part of the story in greater detail. And this does not seem to make sense. (Whereas we regard people as

both speaking and storytelling creatures.) And since speaking in humans is also constituted by capacities inherent to storytelling, it weakens considerably our entitlement to label parrots as regularly 'speaking' and brings them closer to voice-producing machines. (Although in other respects they differ from machines significantly.) Analogously, it is not enough to imagine that we can call machines 'thinking'; the crucial point is to consider *what else* has to make sense concerning the machine, in order that we can talk about it as 'thinking', just as we talk about thinking beings (prototypically people). Turing himself implies a provisional, yet reasonable answer: it is difficult to conceive of something as a thinking entity, unless it is able to speak. It does therefore have to make sense to conceive of machines as speaking (not just voice- or word-producing) entities.

(This notion of thinking—as mental capacity or activity manifesting in speech—is indeed not a particularly controversial one, but still it is rather narrow and should be distinguished from other senses in which we speak of 'thinking'. I will only remark here that thinking in the sense of meditation, or—on the other hand—put broadly, as whatever goes on in the mind, regardless of the outputs, is not that relevant for the following discussion of Turing's proposal. Which is not at all to say they are not relevant for philosophy altogether.)

2. The second option is to treat the meaning of 'thinking' as accomplished and demanding no further scrutiny, and instead to design and apply an empirical procedure diagnosing and revealing thinking entities in practice. So the second—empirical—reading of the opening question is: 'Can a machine that thinks be construed and/or recognized as such? And how?' This approach can be found, e.g. in (Moor 2001), where the TT bears no conceptual argument, but is designed to confirm inductively the hypothesis that AI is possible.

To answer the question, we have to design a viable procedure, the application of which will provide us with predictions with no or only a reasonable amount of mistakes. In usual practice, this requires at least two things. 1) There must be a procedure that *can* be applied in practice and eventually *is* applied and divides the investigated sample in a non-trivial (informative, not self-evident) manner into two groups—individuals that are X and those that are not. 2) The predictions of the procedure should be open to an independent confirmation, meaning that in the end, we have to have a tool enabling us to see whether the test predictions are reliable. Consider a test revealing whether someone has a disease that—in the beginning—has no obvious symptoms. The parameters of the test must be such that it can be carried out in practice, and when it is carried out it should offer non-trivial diagnostic statements dividing individuals into infected and non-infected. The

prediction must be open to confirmation; if not by an independent second test, then at least by the eventual outburst of visible symptoms of the disease. Analogously, an empirical procedure revealing a thinking machine must be performable in practice; it must confer non-trivial information about which tested machines think and which do not and the prediction has to be open to some kind of independent confirmation.

If we proceed in such terms of the empirical, it is, in a sense, easy to deny that machine thinking is an empirical question. For something to be an empirical question, the concepts in terms of which the question is asked must be unequivocally understood in advance. While it had been quite clear what was meant by 'consumption' long before science renamed this condition 'tuberculosis' and began analyzing its intern processes, we do not have an equally clear picture of what it would mean to say that a machine thinks. It is not clear whether empirical and technological experiments have only to tell us more about something we all already know and understand in more or less the same terms. It already points to the uncertainty of whether such a thing as a thinking machine does not happen to be *contradictio in adjecto*.

As it happens, the discussions centered around AI (the possibility of thinking machines) often mix the two approaches together. First, the very nature of the problem excludes, for now, an independent second test, since we are still searching for the first reliable one. Second, some of the criticisms of the TT (including the most famous ones) present peculiar thought experiments that are not performable in practice, at present. There, the discussion seems to stay at the conceptual, speculative level. But this is not adequate either. The question of AI is not a part of everyday conceptual equipment—the presumed principal subject dealt with by the concept-oriented philosophy. AI research, however, uses an *expert* segment of our language that is still being established and changing rapidly, in a direct connection with practical experiments performed by computer scientists, engineers, etc., but often quite disconnected from the way people outside this environment speak. The rules governing the particular use of terms like 'thinking' in this specified context change gradually, more rapidly than in language used by AI-laypeople. This makes the topic unique and unsuitable for the seemingly simple question, 'conceptual or empirical?' answered rather unequivocally in storytelling parrots and tuberculosis. Chomsky (2008) suggests that Turing was aware of the peculiar nature of the debate and knew well that in terms of the everyday language the question, 'can machines think?' was idle and absurd and that only future shifts in our conceptual frameworks would change that. Similarly, once it was just as absurd to think of man flying (because only birds, angels, etc. were capable of flying by themselves). Now it is not as absurd—due to the development of technology, but also due to the

development of the terms in which people think.

This opening consideration sheds some light on the status of Turing's proposal and its major criticisms. What did Turing himself say? His original (Turing 1950, 433) account starts with an explicit admission of bypassing the conceptual issues tied to the question of whether machines can think. Turing's strategy was to avoid the discussion of what 'machine' or 'think' means in common language as something which is a matter for an uninteresting statistical survey. He proposes an *empirical* strategy instead, that tests certain levels of machine sophistication in practice (and could serve this way as an indirect answer to the question).

Nevertheless, there *is* a conceptual level grounding Turing's argument (he, too, makes non-evident conceptual assumptions). First, he wants to separate "thinking" sharply as an intellectual capacity from those human capacities that are in any sense "physical" (Turing 1950, 434). The "imitation game" with which the exposition starts consists of guessing who is a human being/a machine on the basis of their communication outputs. Here, the conceptual context is suggested: to be a thinking being means having an emergent quality testified by the way one speaks. Turing presents a diagnostic procedure, working under specific (empirical) conditions: if the unknown interlocutor passes the test, the interviewer will be entitled, with a solid degree of probability, to assume that she is dealing with a thinking entity. She will be entitled to adopt attitudes analogous to the attitudes which she adopts to thinking beings (humans, by default). Despite the lack of ambition to engage in a conceptual debate, Turing assumes, though *not explicitly*, a notion of thinking that describes a *disposition manifesting* itself in a certain type of expression. For Turing, to think means to be able to behave in a certain way. The approach is dispositional; it does not postulate an entity called "thought" (or the like) occurring within the scrutinized agent, it only describes how particular agents—that we call thinking beings/agents—work in practice.

As usual with terms of this kind, it is rather difficult to distinguish between the manifested and the manifestation. In as far as we apply two different terms here, within our linguistic practice, as *truly different*—speaking and thinking—, the two must not be confused. Otherwise, one of them becomes superfluous; if none of them is superfluous, a distinct meaning must be attached to each of them. Turing's (1950, 434, 436, 445) argument is safe in this respect: he emphasizes repeatedly that his test offers a *criterion* for thinking. Roughly speaking, although thinking is expected, for the sake of the argument, to be captured—in both necessary and sufficient manner—as an ability to behave/express oneself in a certain way, it is *not*claimed to be one and the same as behaving/expressing oneself in certain way. Proud-

foot (2005), e.g., therefore understands the TT as externalist, not plainly behaviorist. This is perhaps not completely adequate, since for Turing the determining factor is not the methodological importance of outward criteria (which is, for instance, more or less, the later Wittgenstein's case and the reason for his reserve towards unexpressed mental contents) but his decision to focus on the expression of intellectual capacities as separate from physical ones. (To be fair, Turing never claims thinking to be this or that; but he—clearly enough—assumes something of such an "externalist" nature. Unlike the later Wittgenstein who admits the variety of ways we speak of 'thinking'—sometimes discontinuous and sometimes incompatible—Turing seems to be close to the early, Tractarian Wittgenstein for whom applying signs in propositions is probably crucial for thinking (Wittgenstein 1961, 3.5, 3.11).)

The notion of *criteria* is both important and interesting. Very cautiously, Turing avoids the dangerous talk of necessary or sufficient conditions, since criteria cannot be identified with either of these. Lyons' (1974) subtle analysis of the concept of 'criterion' proposes to understand criteria as "any *important standards* by which we judge that something is an X, and hence they may be related either empirically or *a priori* with X"; that is, by introducing his proposal as a suggested criterion Turing seems to avoid the need to stand on either part of the boundary between empirical and conceptual (see also Chomsky 2008). However, there are other difficulties tied to criteria; I will expose them in closer detail later.

The nature of the criticisms of Turing is often mixed in the context of our opening distinction, too. The most famous counterargument is presented by Searle (1980). His Chinese room argument attempts to demonstrate that operating with input symbols and generating outputs according to pre-set rules is—to a great extent—independent of thinking. Searle's *thought experiment* is designed to illuminate thinking as an intentional aware activity, emerging from biological grounds; it thus makes no sense to talk about thinking, unless a living being that thinks is concerned. Yet Searle's argument can be read empirically as well: as if he suggested that it would be the *result* of the Chinese room experiment (and in the *moment* of the result) what would refute the validity of Turing's proposal: "If such and such conditions could be met (and I do not see why not), your procedure would prove to fail because it would "reveal" a thinking machine while there would be no such thing." Meaning that under such and such conditions—which Searle perhaps does not consider impossible—the TT will prove to be faulty: we will get a result that can be independently checked as false.

A similar argument is provided by Block's (1978, 1981) idea of the Aunt Bubbles Machine that could bypass the TT by means of an explicit man-

ual displaying all the possible ramifications of the conversation. The target of Block's attack is behaviorism in the philosophy of mind (to be proved as mistaken using the TT as the focal example) in the first place, not the TT itself. Block's is a thought experiment of the hybrid empirical-hypothetical nature, too: it suggests that if we were able to equip someone/something with such a manual, the TT would give these counterintuitive results. Just as Searle, he assumes the meaning of 'thinking' to be known, unproblematic and possible to check. An interesting objection, however, can be made concerning the idea of the manual: and it is debatable if we will ever be able to create such a thing, given that Chomsky (1980) points out that the expressive freedom/ability open to the competent human speaker is inexhaustible or unlimited.

Both Searle and Block seem to accept a paradoxical proviso. The chatbot training, according to their counterexamples, takes shape of the endeavor to *predict* every possible conversational situation and determine fully the chatbot's necessary algorithmic equipment. However, we expect from a thinking being, based on the human model, some degree of unpredictability in behavior. (Some recent scientific accounts of free will (e.g. Brembs 2011) propose that it is reasonable to model free agents as beings acting in an *unpredictable* way under conditions that are still the same. And the concept of the thinking being, in the ordinary "human" sense, is after all closely linked to the concept of the autonomous agent with free will.) Turing, on the other hand, worries much less with the need to predict every step of the machine. He points, in reply to the putative "Lady Lovelace Objection", that the need to provide a full computational background for the performance of the machine is no obstacle for considering it as thinking. Since humans cannot trace the whole complexity of the computations, machine performances are somewhat emergent from the computational substrate. That machine thinking is fully computationally determined does not prevent it from performing in a way that is surprising and novel for human observers, including those familiar with computational equipment like Turing himself (Turing 1950, 450f).

Stanisław Lem (1996), in an elaboration of Searle's argument, focused on the conceptual level of the problem. He points at co-extension between thinking and the ability to pass the TT. Based on the "imitation game", it is the ability to imitate *human* expression that is crucial for Turing. This ability stands proxy for the ability of *human* thinking. Lem accepts Searle's insight that what we call "thinking" is inextricably tied to the human biological/physiological basis and he asks: Why should we expect a machine capable of autonomous thinking to express this ability the very same way humans express *their* thinking ability? If artificial thinking is an expression of a dis-

position emerging from the particular *mechanical/computational* basis, then a particular, machine-specific expression can be expected, too. A thinking machine could then fail in the TT. The core of the problem is the *conceptual* identification of "thinking" (without qualifications) with "human thinking" or, to put it more precisely, with what manifests the way human thinking manifests itself (see also Millar 1973). Unlike Searle or Block, Lem does not fantasize about any possible (?) empirical procedure to undermine the TT, but tries to show that the concept of thinking operating on the TT is too narrow and can lead to a dead end. In his idea of the "uncanny valley", Mori (1970) points out that the closer the machine's expression is to the human, the more dissatisfactory the result will be. Neither Mori nor Lem thinks that the goal of imitation cannot be reached, but they both dissuade computer specialists from it (though Mori's argument has been criticized as empirically inaccurate and anthropomorphism in HRI (Human-Robot Interaction) research has been defended as natural and not necessarily wrong—see e.g. Zlotowski et al. 2013).

Lem's case is interesting because Turing anticipates briefly an answer to his points: according to Turing, it is not important if it is actually this ersatz capacity to imitate human (linguistic) behavior which the TT reveals—if we are able to prepare the machine for successful performance in the TT, that is enough and nothing more need worry us. Likewise, though machines could be expected to display a particular, different kind of machine thinking, if they are able to pass the TT (display a human kind of thinking), it need not worry us (Turing 1950, 435).

The TT fulfills the condition that it can be and actually is applied in practice. The empirical fallibility of the TT is most often demonstrated in two main directions—as being both too narrow and too wide (Block 1981). First, it is too narrow and fails to detect all cases of thinking. Animals are often considered as having thought too; the notion, "intelligence," is sometimes applied in their case as well. Though the opinions about animal thinking or intelligence vary and the issue is difficult in general, the TT verdict would be unequivocal here: animals do *not* think. But that is far from obvious. Also, people are not a homogeneous group; various speech (and/or also intellectual) capacities are distributed among the interviewers as well as among the interlocutors. This is also one of the reasons why the Loebner Prize competitors have already long ago found it necessary to incorporate simulations of "artificial stupidity" (Liebowitz 1989) into their experiments as well. However, this has not enabled them to reach the final goal either; even though Turing himself saw making mistakes and displaying "frivolous" behavioral varieties only as questions of proper programming and storage capacities (Turing 1950, 447ff).

Second, the test is too wide and—under certain conditions—it can be passed even when no thinking is involved. The famous story of Weizenbaum's (1966) ELIZA program and its temporary success in the TT is well known. Under these circumstances it is easy to object to Turing on the grounds that the credibility of the test is doubtful when a chatbot succeeds in a test that has already failed. Some suggest (e.g. Whitby 1996) that the case of ELIZA and the reverse cases (such as the Shakespeare accident) only document that Turing's proposal has led AI enterprises into a blind alley.

If we stay at the empirical level, the fallibility of the test is not a grave problem. It is true that Turing's proposed procedure seems to be weak and erroneous. But the history of medicine (e.g.) is full of examples of weak diagnostic criteria for various diseases that, however, have ultimately been gradually made more precise. If this was the only problem for Turing and for those trying to pass the TT, it could be justly expected to be solved in time. I think more serious problems lie at the conceptual level; in the next section I will present some of them, grounded in the later-Wittgensteinian perspective. Strictly speaking, philosophy has nothing to do with predicting whether, how or when a chatbot able to pass the TT (either the original or an advanced version) will occur. But it can judge, when/if it arises, whether and in what sense we are entitled to apply the concept "thinking" to the successful software.

3. Two Wittgensteinian Objections: The Concept of Language

Some of Turing's ideas on account of AI have been subject to criticism by Ludwig Wittgenstein, mostly due to the notion of thinking employed. Despite Turing's unwillingness to participate in conceptual debates, he adopts a certain notion of thinking and takes it for granted in his empirical proposal. I will show here, using the example of two authors reflecting on Turing from a rather strong Wittgensteinian position, that the implicit concept of speaking (language) that he uses is superficial. If we expect thinking to be certified by speaking, we must consider the full range of what 'speaking' means. From the Wittgensteinian point of view on language, many AI arguments suffer here from a simplistic focus.

Carlo Penco (2012) thus reflects upon the putative Turing-Searle debate. Turing, in his pivotal 1950 text, limited himself very cautiously to designing the test procedure alone. Searle, however, acts as if Turing was promoting a complex conception of mind, understanding and language. Turing—*sensu* Searle—proceeds as if intelligence lies only in the capacity of manipulating symbols (language signs). Searle's thought experiment aims at showing that this mistaken *concept* of intelligence would lead the TT to fail in *practice*.

According to Searle, understanding the underlying language capacity

(that affirms thinking) cannot be exhausted only through manipulating symbols based on a manual. This is possible even without understanding the meaning of all this: by the "man inside the Chinese room", for instance. So though we see the people who do understand and speak only as symbol-producing creatures, they would not be able to do so if there were no organic process of internal thought, hidden behind the symbol-manipulation facade.

Penco's account agrees with Searle that the sole symbol-manipulation, as Searle understands it, is not enough. If this was the only thing required, the scenario of the possible test failure, vis-à-vis the Chinese room, would be relevant. But even though such an account of understanding is, in reality, crudely narrowing, the way to its enrichment does *not* go through capturing the inside of the mind. Wittgenstein, to whom Penco refers, points out that actual (observable) language games that are played cannot be explained through an ontology of the "inner" (mental contents, emotions etc.), but rather by further explanatory language games within which "mental contents" and "emotions" are concepts playing limited parts (2009, §§654–6). We understand what people say, not by means of an investigation (scientific?) of their interior, but by means of further talking. Penco skips Wittgenstein's admission that "the mental", if understood properly, contributes to understanding; rather he points out that the language games are made available by exposing them as embedded within a *context* and by explaining the context, if needed. As in real conversation the TT should involve, not only questions about factual information (consider here the curious Shakespeare accident reported by Halpern 2006), but also, and first and foremost, statements and questions with *indexical* terms that even in a seemingly simple shape represent a major challenge: "This is a nice colour", "This man is someone familiar to you, isn't he?" or "You cannot trifle with me" [whereas with someone else you can?] and so on.

A competent reaction to these baits cannot consist of symbol-manipulating capacity *alone*. According to Penco, the capacity to manipulate symbols *in context* is needed; and if Searle sees the linguistic ability as consisting of symbol-manipulation alone and apart from context, that is incorrect (in this argument, Penco relies on Diego Marconi's (1997) analyses). This is rather a Wittgensteinian than a Searlian concept of understanding; being naturally *situated* into contextual practice (and acting as a context-situated agent) represents the main challenge for TT competitors, not the necessity to implant the organic-like understanding-procedures into the "mind" of the machine. Turing himself does not thematize the role of the context in his implicit account of what thinking is; his proposal is thereby not refuted, it has only to be broadened. What Penco does is rather a *precision*: he points out that the

human interviewers implicitly and naturally assume that their interlocutors are able to emit symbols displaying such context orientation as well. It is appropriate to say here that "context" means not only incorporating the indexical expressions (conditioned by sense receptors), but also understanding the importance of various social roles the participants of the conversation can play, etc. Though, for the definition of the test, this is only a marginal addendum, it may represent substantive difficulties for the "training" practice. The point is to *admit* that the way the test works presumes our actually used *concepts* of 'understanding', 'speaking' and 'thinking,' which include a rich contextual content, and to assess the requirements, chances, etc. under the light shed by this admission. Interestingly, this is an objection not so much to Turing, as perhaps to Searle's ideas about what is wrong with the TT.

Another—partly related—Wittgensteinian comment is made by Stuart Shanker (1998). Unlike many commentators of Turing's proposal, Shanker explores Turing's pre-1950 works in light of which he interprets the Turingian notion of AI. He exposes the root of Turing's proposal to be his (Turing 1936) account of *calculation*, where the observed ability is explained in terms of its unobserved (inner) causal conditions. Thus what makes the machine capable of calculating (and later, thinking) is its particular inner equipment. I am not sure to what extent Shanker's interpretation can be extrapolated to the later proposal of the TT, as its ambition to posit anything inside the tested entities is very modest, if any.

However, Shanker sees Turing's main problem to be his view of following a rule. In order to make the machine capable of acting according to rules, the rules have to be reduced to simple sub-rules of a purely computable ("mechanical", as Shanker puts it) nature that can be followed by a machine without the need to interpret them. However, though Wittgenstein himself speaks of a *mechanical* rule-following as underlying the human speech capacity, it is not the same as to causally pre-determine every step the machine makes—to create an algorithm. What makes a skill—such as calculating or thinking—a real skill is its *normativity*. Mechanical rule-following means that the agents do not need to reflect upon the respective rule when they follow it—they have mastered a practical skill. They should, however, be able to correct or explain their behavior with reference to a rule (or at least to attempt it, understanding that in certain situations they are legitimately requested to do so). Speech that manifests thinking is speech consisting of rule-following moves; but this normative procedure also involves instructions, imitations, sanctions, successful and failed attempts, explanations, etc. We admit the status of a rule-following agent to *no* entity unable to link corrective or explanatory moves to its (seeming?) speech statements, or to display a believable willingness (acceptance of, not a universal, but a

case-by-case necessity) to do so. The question follows then: Is a machine that claims some intelligence for itself an entity that can be reasonably demanded and expected to do such a thing?

Shanker's critique of Turing's machines seems to omit the fact that the individual performances of the machine do not need to be completely determined; Turing himself admits random elements in the computer's algorithms (Turing 1950, 438, e.g.). This does not, however, blunt the edge of the critique: a stochastic machine is perhaps closer to unpredictable human performances, but cannot be expected to act as a normative agent either. If, according to Wittgenstein, Turing's idea of machines capable of doing the same things as people do (calculating or thinking in the human sense) lies in the machines being equipped with a specific internal (be it fully causally determined or stochastic) background, it is simply mistaken. This does not amount to normative behavior.

In Shanker's terms, Wittgenstein's critique carries no empirical moment. Wittgenstein was probably not familiar with the idea of the TT (that appeared only shortly before his death); thus we can only conjecture that he would have seen the main problem of the *test*, not in its possible empirical fallibility, but in the fact that Turing misunderstood what 'thinking' means. Thinking, especially when diagnosed on the basis of speech, can be attributed only to where normative attitudes are present. Certainly, "normative attitudes" point to a rather old-fashioned, Kantian-like or Sellarsian concept of rule following, where normative agents are expected to be able to adopt reflexive attitudes to rules, oppose them or question them (talk about them, in general). They have to be able to give their reasons and ask others for theirs. When Shanker—in Wittgensteinian mode—speaks of following the rules, together with the practice of violating or bypassing them (being right and being wrong), he also stays within this framework. On the other hand, artificial entities like the present intelligent systems (computers running search engines, for instance) can be taught to follow certain rules and failure to follow those rules is not exceptional. The point of Shanker's critique is that unless it makes sense to speak of machines following and breaking rules in the *former* sense, there can be no thinking machines, regardless of any successful results from the TT.

On the other hand, this is not completely fatal for the idea of the test. 1) There are also relevant accounts that try to construe thinking entities as capable of normative agency on a computational basis; see, for instance, Dennett's (1996) analyses of intentionality, including the "derived" intentionality in robots. (What remains here is to assess whether this notion is closer to AI-laypeople's manner of talking about "thinking" than the normativist one.) 2) If every explanation or strategy focusing on the machine's

causal/stochastic *intern* is cautiously removed, it may be useful for searching non-human thinking entities. One need only admit that the test works (or should work) in the way that the interlocutors implicitly expect *normative* behavior from a thinking entity. The reading of the TT has to be only enriched with the *normative* level of the studied phenomenon. I think it is Turing's emphasis on the test's being a criterion, that leads him to substantial (though often overlooked) problems; more than his omission to mention the role of context, and perhaps even more than his negligence of the normative dimension of genuine thinking.

4. The Problem of Criteria and the Life of Concepts

The most serious problem for a literal understanding of Turing's proposal is, I believe, represented by Turing's emphasis on the notion of criteria. He admits (Turing 1950, 435, the end of Section 2) that he knows no better criterion for a thinking machine than its success in the test designed by him. It is thus *a* criterion both sufficient for him and—only as far as he has nothing better to hand—necessary. This provisory claim amounted later, after Turing, to the practice claiming the success in the TT as *the* criterion for thinking. Let us look at the question of criteria in closer detail. (Like Penco and Shanker, I will make use of some of Wittgenstein's arguments here too, but I will draw some consequences from his more general account of thinking and speaking, originally unrelated to his debates with Turing.)

"Criteria" need not be necessary or sufficient conditions of a phenomenon subjected to empirical scrutiny. If X is a set of criteria for Y, then if it makes sense to state X, we are entitled to reasonably infer Y as well. A criterion for a phenomenon belongs here to the *notion* of this phenomenon, as its contingent *a priori* (Rorty 1989) *constituent*. Lyon (1974) states that a criterion stands in an analytical relation to the thing itself, being a concept that is a constitutive part of the other concept. He offers here a little confusing example of having pips, which is *one* of the criteria for being a lemon (certainly, the two concepts cannot be confused—they are not predicated *promiscue*). From this point, three major problems for Turing arise.

1) Turing's case is a peculiar criterion proposal. Its first problem is that in other phenomena, there is usually not *only one* criterion; that is, we do not usually predicate things based on a unique *and* exclusive criterion (not expected to imply anything else), but usually a diagnostic *set* of criteria. On the other hand, *the TT is meant to be just such a unique coextensive criterion*. And what is more important, even if we could work with an isolated criterion, the search for the occurrence of a criterion *must not* be confused with the active endeavor to carry out the criterial situation/phenomenon/event. "Great artists always have a lot of women, so if he can have a lot of women

that makes him a great artist," is a funny example of such a mistake (from *Lucky Jim* by Kingsley Amis). One can easily commit the same mistake if he/she i) identifies thinking exclusively with passing the TT, and ii) focuses actively on how to make his/her software pass the test, instead of "making it think" (which is naturally too vague a task, if put this way). In both respects, Turing stands midway in committing the mistake, as it were: though he does not identify machine thinking with passing the TT, he suggests focusing on the TT as the only meaningful way of dealing with the question of machine thinking; and though he does not focus actively on creating software, he suggests that doing so is no mistake, and the safe way. But what use can we make of a unique independent criterion? Can such a thing be found anywhere, outside the TT context—something used and useful for exclusively detecting anything different from itself? Hardly—and in this sense the endeavors to bring about the TT "criterion" are well justified. If in machines, "thinking" is really nothing more than successfully passing the TT, why not try to make it do so? Either a criterion is not unique and coextensive, or it cannot be distinguished from the thing itself, and it then makes no sense to speak of it as if it was something other than the thing itself.

Thus it appears that the relationship between thinking and speaking, even in humans, whereof Turing took the model for his test proposal, *is not that of there being a unique coextensive criterion*. One possibility is that there is not just one criterion but, as standard, a set of several criteria. That speaking may be only one of them (so that other criteria have not been taken into account) could explain the cases of empirical fallibility of the TT, identified many times. This includes the cases when the interviewers regarded humans as machines (or *vice versa*), or the fact that the TT cannot capture thinking in other (probably) intelligent beings, like animals, etc. These empirical flaws can be expected to undergo correction; the test will improve. The more serious problem is that there is no partly, or wholly independent confirmation tool. Our diagnostic procedures for various diseases are in the end checked at least by the outbreak (or non-outbreak) of the disease; but the TT prediction cannot be checked.

Yet the uneasiness felt in cases such as the success of ELIZA does not come from applying an independent, second standard with dissenting results. It is not that from a set of criteria only some were met. Rather it was, from the beginning, much doubted that something like ELIZA could be reasonably called "intelligent" or "thinking". The cases of human interlocutors evaluated as machines, thus, show the TT as a fallible (and corrigible) empirical tool; but the ELIZA case shows that the TT fails as a criterion. If it was a genuine criterion—provided that a unique criterion could work, which it most likely cannot—nobody would be inclined to question the re-

sults. The motivation to demonstrate that the TT must have been wrong in ELIZA's case is based on our conceptual intuitions. ELIZA was eventually unable to pass the test repeatedly under standardized conditions (when interviewers knew that their interlocutor might be a machine). But if someone/something regularly passed, the very nature of the test should make us admit the occurrence of AI. Why are people then still reluctant? It is our presumptions about thinking that hinders us (*petitio principii*, to be true). Some such observation probably lurks behind the simile Wittgenstein (2009, 174) uses to explain why we are so unwilling to assign thinking sometimes even to animals: "If a concept refers to a character of human handwriting, it has no application to beings that do not write." We are just not willing to let ourselves be convinced by speech production, accompanied by something else or not, that a machine thinks. Under this consideration it seems that *speaking is not one of the several criteria for thinking (instead of being the criterion), but rather no criterion at all.* The way we ascribe thinking has not much to do with criteria.

Let us look more closely at the example of humans. Somebody may object that here Turing's proposal works: we assume humans to be thinking beings, though we have nothing as evidence except the criterion of their external—mostly speech—manifestations. But what looks like a unique exclusive criterion, does not play this role in reality. A criterion is something on the basis of which, if stated, we are entitled to state something else; here it would mean, *on the basis of other people speaking*, we would *conjecture* about their status as thinking beings. But this does not work in such a way; this pattern of criteria use occurs when we are searching for something, and when we are in doubt (where the searched phenomenon is recognized properly by some—perhaps many—but can also be misrecognized). Wittgenstein (2009) asserts that criteria are at play in the case of person- and moment-specific moods, ideas or dispositions (see, e.g. §§269, 579f); whilst on the other hand we cannot state as a result of reasoning that someone is endowed with thought, because that is a matter of our attitude to him/her (Part II, iv). Winch (1980–1981) interprets Wittgenstein's remarks to be that we are, from the very beginning, "set up" (*eingestellt*) to adopt the "attitude to a soul" towards other humans; that the others do think thus, serves us as an *argument*, or *premise*, in discussions, and we *do not conclude* that on the basis of hearing their speech. Neither do we have to, nor are we able to perform such an inference. That others think is one of the central nodes in our *conceptual* equipment. In a way, *we ascribe speaking on the basis of thinking* (i.e. we interpret people's emitting vocal outputs as structured and meaningful linguistic behavior, because our foundational attitude to them is an attitude to thinking that beings are capable of such a specific pattern of behavior,

unlike parrots), *rather than otherwise*. But not even that is an inference (cf. Wittgenstein 2009, §357: I do not say X thinks on the basis of observing X's behavior, but claiming that X thinks makes sense only given X's behavior).

Unfortunately, the TT allows space to assume that such an inference from speech to thought can or should be made deliberately and step-by-step. Parrots can be taught to pronounce many words and sentences, and sometimes to apply them on seemingly suitable occasions. But not even then do we ascribe thought to them in a human sense. Thus, even though speaking when connected to distinct stimuli *means* thinking, this is not an inference we make. In our discursive practices, "thinking" does not play the role of something that has to be carefully judged and only then attributed to the scrutinized subject; we are not accustomed to working with the concept in this way. And since we do not perform a step-by-step inference, there is also no room for stopping halfway. The strong dispositional concept of thinking allows for its gradual nature, but our usage contradicts that: we do not say that an eloquent intellectual is a thinking being "more" than a taciturn peasant (though we may be tempted to say he/she perhaps thinks "better"). One is either a thinking being as humans generally are, or is not at all; the changes and learning stages undergone through life are not taken into account in this intuition. The uncertain status of animals is based on our mixed discursive dealings with them. It is not due to their inconclusive performance in an exam (of the kind of the TT). Rather, people deal with animals as with thinking agents in some contexts, while as with inanimate property in other contexts. Creating more and more accomplished chatbots is quite similar to the transitional stages in the life of small children who do not speak properly yet. But neither are such children denied "thinking" (the status of a thinking being); despite Wittgenstein's (2009, §30) skepticism we prefer interpreting their performances as 'thinking not yet developed' rather than 'thinking absent'. For some reasons, we do not apply this conceptual strategy to machines.

Again, this is not an argument against the thesis that amazing results can be achieved this way in practice through technology. Turing proposes thinking as an emergent quality (similarly to Dennett, much later) arising from computable grounds. In his second, more thorough answer to "Lady Lovelace's Objection," (Turing 1950, 455ff) he argues that the effect of surprise or novelty can be reached, e.g. as a super-critical output reacting to a sub-critical input added to the sub-critical mass of programming equipment. It is only a question of having enough sophisticated programming to achieve this. It would be proper to say, in favor of Turing, that the origin of human thinking was gradual too, and no fully-fledged thinking beings appeared overnight. The substantial obstacle he faces is not the empirical im-

possibility, but the conceptual classification of the—prospective—machine successful in the TT. For our present concept of (human) thinking is largely independent of the probable empirical development of thinking.

The necessary gradual development in the field of AI makes machine "thinking" more similar to a different family of concepts, the use of which is structured in a particular way. If we are to start calling a machine "thinking" on the basis (and from the moment) of its victory in the battle with the TT, we adopt a standpoint analogous to the standpoint adopted to a student who has successfully gone through graduate study, passed all exams and defended his/her thesis in open discussion, by virtue of which he/she can be henceforth called "doctor". Here, too, a case of a unique and coextensive criterion is concerned—and we see that passing all the required exams, etc. amounts to being a Ph.D. But this is exactly the problem: we do not distinguish between the two. 'Doctor' is here nothing but a kind of (honorary) *title abbreviated* for a person who has successfully negotiated his/her way through the graduate study course. It makes no sense to ask whether she "really" is a doctor. 'Doctor' is an inferred title; 'thinking being', however, is not. If it was, nobody could contemplate the idea of denying the successful chatbot the title, 'thinking entity'. Yet if it succeeds, as in the moment when the interviewers see that the interlocutor is a machine, the uncanny valley effect (triggered by our mostly unconscious linguistic intuitions) will—in many people—activate the reaction that says "this cannot be genuine thinking". For Wittgenstein (2009), the "reason" for this incredulity is that we have learnt language in which "we only say of a human being and what is like one that it thinks", whereas acquiring the status of (sufficiently) "being like a human" is difficult since this is not an empirical statement (§§359f). Consider the original imitation game: a successful female player does not become a "man" or someone who thinks the way men think—despite the good performance. At the moment of disclosure (at the latest) she ends up with the title of a good man-imitator, based on her ability to act, for some time, *as if* there were no gender difference. Similarly, the admirable skill and mastery of computer scientists may clash with the observers' depressing stubbornness at suspecting the machine to be only capable of performing *as if* it thought.

The two patterns—set-up ("Einstellung", *sensu* Winch) vs. title—of operating with reference to the term, "thinking," differ in the conditions of their application and in their positions in our inference practices, etc. To say that machines are not on the putative list of subjects admissible as "thinking" (seemingly implied by Wittgenstein's remark) is not just a plain fact, but it is embedded in our discursive practice. The problem with thinking is not just that most people do not speak of thinking machines (Turing's Gallup argu-

ment). How most people speak is a pattern established through and as a practice that *works*. To contradict it means not just to oppose the speech habit of most people, but to tear apart a rich context of *inter-conceptual links*. The well-worked practice of using the word 'thinking' also implies that thinking beings are capable of thinking *about* something, have beliefs and opinions (often about other people), act—including speaking, if they are capable of speech—on their own behalf, etc. It is preposterous to demand that a machine be recognized as speaking, if it is not also prepared to demonstrate these interlinked capacities. Wittgenstein (2009, §361) also points (in not quite a clear manner) to another contextual direction. Whereas people (and, in a different sense, animals too) count as the agents *who* think; in the case of disputable entities like machines and chairs, we are still tempted to ask such questions as—"*Where* do they think?" which can only result in nonsense.

Human thinking is a result of a highly complex evolutionary development, the subtleties of which we have not yet fully understood. Also each person in his/her life thinks and gets accustomed to thinking long before he/she comes to asking the question whether others (or he/she, herself/himself) think. Some recent research in computer science, e.g. in evolutionary robotics, has made interesting contributions to our knowledge of the evolution of thought (see, e.g. Rohde 2009). Among the more naive and less serious AI enterprisers, the exemplar of which is offered by the Loebner Prize competitors, we are witnessing attempts to avoid any slow development from unaware origins: the developers try to achieve status *via* focusing intentionally on the speaking criterion only. But we saw that i) the reversal of the status-criterion hierarchy in the end amounts to the *identification* of the two, and ii) that speaking is probably neither *the* criterion nor *a* criterion for thinking at all. The success in the TT therefore means that the programmer has created a chatbot able to succeed in the TT; to assume more means to ignore the subtle relationship between the non-identical semantics (semantic fields) of the words 'think' and 'speak'.

Unlike the other two Wittgensteinian objections presented by Penco and Shanker, these problems concerning criteria can be grave. Turing does not seem to understand that a unique coextensive criterion—a prominent example of which is supposed to be human speech corroborating human thinking—perhaps cannot work at all, and that thinking, unlike genuine criteria, definitely does not stand as a premise in an intentional and aware argument (inference). And both Turing and his followers do not seem to understand that one cannot achieve X using a strategy that consists of intentionally yielding the criteria for X (this has also been remarked upon by Whitby 1996). Because the time scope of the TT is limited, this strategy has the effect of

building a restricted skill, somewhat like when someone needs to pretend to be somebody who he/she really is not for 30 minutes, and then relies on producing the most uncharacteristic gestures, manners, etc. The test thus reveals a certain level of programming sophistry (that can be used fruitfully in practice, to be sure), just as differently designed (easier or harder) tests reveal different levels of programming sophistry.

Turing, with his declared lack of interest in how people commonly speak/spoke, would have readily embraced the concept of "thinking" in a different sense. After all, there is a lot of hard work behind passing the TT, as there is in becoming a Ph.D. For him, it is irrelevant whether speakers are reluctant to assigning the title 'thinking'; the success in the TT is achievement enough. As we have seen, Turing in fact attempts to restrict the term (to separable, purely intellectual capacities). However he seems to ignore the fact that if one just decides to use a term in such a differently defined way, its use in practice (going beyond the reach of one person's decision) cannot be expected as a matter-of-course. *He neglects the "natural" dynamics of the development of our semantics*. The meanings of our words and phrases no doubt develop and change, but the result is not given by one individual's stipulation. The same applies to such "thinking" conditioned by explicit criteria (within which speaking plays a principal part); it may not be ruled "logically" or once and for all; but one cannot simply enforce such a shift.

This third mistake committed by the TT is partly independent (and, somewhat curiously, brings hope that someday the problems with criteria can be bypassed). Even if a long-term, non-restricted success in the TT occurred, it could not render machines "thinking". This can happen, as it seems, only through a shift in our conceptual frameworks. Language is a living, developing organism, reacting to its users' intentional attempts at changes. Although nobody can warrant that the reaction will correspond to the innovator's intention exactly. So there ultimately *could* be a chance to meet (to speak of) "thinking machines" someday, but this seems not to depend *only* on the computer engineers' and scientists' endeavors. Their achievements would have to be supplemented someday by the enrichment of the conceptual dialectics of our language, encouraged by them; until that day any success in the TT will probably remain only a "success in the TT". Since this is no more than a conceptual precaution, participants in the TT need not worry. It is probable that someday somebody will succeed in the test. It is not a question for philosophy to ask whether the TT can be passed, how to achieve it, or when it will happen. One indeed *can* ask a philosopher, "Can there be artificial intelligence/thinking?" But he/she should make sure he/she knows what he/she is asking. One important thing a philosopher can determine is whether his/her concepts do not contradict themselves, just by

asking this question. For the time being, regarding present language usage and the semantic fields of 'thinking', it seems probable that some contradiction occurs here. And it will remain so until the notion changes, which requires time and a gradual (natural) course of development. A conceptual shift among pioneers of discipline is probably needed first (as in the case of people capable of flying), and technology can help disseminate it among a wider range of people. (Certainly it is easier to disseminate that it is not an absurd idea that people can fly, than that machines can think—the demonstration potentials vary significantly here.)

In evaluating the proposals and achievements in the field of AI, the analyses (inspired by Wittgenstein) I have introduced here can help us recognize the most common shortcomings characteristic of naive readings of the TT, and understand why, or in what respect (to what extent) success in the TT does not easily qualify it as a thinking machine. Here are a few of the most important:

1) Since human linguistic behavior cannot be considered a symbol-manipulation analyzed into simple, purely mechanical/computable and fully explicit rules (at least we are very far from identifying any underlying neurological mechanism, supporting in a traceable mechanistic way our rule-governed linguistic behavior), it is uncertain what guidance this could offer for creating a thinking machine.

2) Since in humans, the boundary between thinking and non-thinking is not set by a difference in inner determining mechanisms, but normatively (though we can otherwise have good reasons to assume that there is a causal background for linguistic behavior of thinking human beings), it is uncertain whether such a background could/should be searched for when creating a thinking machine.

3) The focus on bringing about a particular criterion, through which the status of thinking/intelligence is allegedly achieved, can easily establish a rather restricted skill with a twisted aim. Criteria may tell us interesting and useful things, but preferably if observed and searched for, not intentionally executed.

4) The terms of the test, if any, should thus be such as to disallow the contestant to prepare for fully specified conditions of the test procedure known in advance; consider here the more demanding versions of the TT for which the preparation would be much more difficult, such as Harnad's (1991) or Schweizer's (1998).

5) Patience on the part of AI specialists seems to be needed; changes to how the concepts of our language work follow only slowly after

scientific and technological achievements (the status of "thinking machine" cannot be commanded out immediately).

6) Until that day, it would be perhaps better not to try to exploit speaking as a criterion for thinking (since in humans—with whose thinking and speaking we are best acquainted—it does not work this way).

In summation, when thinking of and criticizing the idea of the TT, we have to distinguish between empirical and conceptual levels. The TT has been shown to be fallible as an empirical procedure; however imperfect empirical tools can usually be improved, if we are clear about our terms. But we cannot take the TT as a fully-fledged conceptual criterion of thinking either; i) because the present semantics of our language, as we speak it, more or less prevents machines from qualifying as "thinking", whatever the result of the TT; and ii) because speaking does not stand as a premise for an inference to thinking, let alone serve as an effectible proxy for the goal itself. On the other hand, Turing may well have also admitted to the first objection against his test proposal, claiming that he was not interested in Gallup polls. Also his subsumed concept of thinking is not substantially mistaken and has respectable philosophical parallels, both as a dispositional (thinking as it manifests itself in rational linguistic behavior) and an emergent (thinking as a novel quality emerging from a computable substrate) concept.

Acknowledgements

This study was supported within the project of Education for Competitiveness Operational Programme (OPVK), Research Centre for Theory and History of Science (Výzkumné centrum pro teorii a dějiny vědy), registration No. CZ.1.07/2.3.00/20.0138, co-financed by the European Social Fund and the state budget of the Czech Republic. I would also like to thank Radek Schuster for encouraging me to deal with this topic and to the anonymous referees for their useful comments to previous versions of the paper.

Bibliography

Block, N. (1978). Troubles with functionalism, *in* C. W. Savage (ed.), *Minnesota Studies in Philosophy of Science, IX*, Minnesota Press, Minneapolis, pp. 260–325.

Block, N. (1981). Psychologism and behaviorism, *Philosophical Review* **90**: 5–43.

Brembs, B. (2011). Towards a scientific concept of free will as a biological trait: Spontaneous actions and decision-making in invertebrates, *Proceedings of the Royal Society B* **278**: 930–939.

Chomsky, N. (1980). *Rules and Representations*, Columbia Press, New York.

Chomsky, N. (2008). Turing on the "imitation game", *in* R. Epstein, G. Roberts and G. Beber (eds), *Parsing the Turing Test. Philosophical and Methodological Issues in the Quest for the Thinking Computer*, Springer Science, Dordrecht, pp. 103–106.

Dennett, D. (1996). *Kinds of Minds. Towards an Understanding of Consciousness*, Basic Books, New York.

Halpern, M. (2006). The trouble with the Turing test, *The New Atlantis* **11**: 42–63.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem, *Mind and Machines* **1**: 43–54.

Lem, S. (1996). *Tajemnica chińskieho pokoju*, TAiWPN Universitas, Krakow.

Liebowitz, J. (1989). If there is artificial intelligence, is there such a thing as artificial stupidity?, *ACM SIGART Bulletin* **109**: 26–28.

Lyon, A. (1974). Criteria and evidence, *Mind* **83**: 211–227.

Marconi, D. (1997). *Lexical Competence*, MIT Press, Cambridge, MA.

Millar, P. H. (1973). On the point of the imitation game, *Mind* **82**: 595–597.

Moor, J. H. (2001). The status and future of the Turing test, *Mind and Machines* **11**: 73–93.

Mori, M. (1970). The uncanny valley, *Energy* **7**: 33–35. Translated by K. F. MacDorman and T. Minato.

Penco, C. (2012). Updating the Turing test: Wittgenstein, Turing and symbol manipulation,, *Open Journal of Philosophy* **2**: 189–194.

Proudfoot, D. (2005). A new interpretation of the Turing test, *The Rutherford Journal: The New Zealand Journal for the History and Philosophy of Science and Technology* **1**.
URL: *http://rutherfordjournal.org/article010113.html*

Rohde, M. (2009). *Enaction, Embodiment, Evolutionary Robotics. Simulation Models for a Post-Cognitivist Science of Mind*, Atlantis Press, Amsterdam.

Rorty, R. (1989). *Contingency, Irony and Solidarity*, Cambridge University Press, Cambridge.

Saygin, A. P., Cicekli, I. and Akman, V. (2000). Turing test: 50 years later, *Mind and Machines* **10**: 463–518.

Schweizer, P. (1998). The truly total Turing test, *Mind and Machines* **8**: 263–272.

Searle, J. (1980). Minds, brains and programs, *Behavioral and Brain Sciences* **3**: 417–457.

Shanker, S. (1998). *Wittgenstein's Remarks on the Foundations of AI*, Routledge, London.

Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society* **42**: 230–265.

Turing, A. (1950). Computing machinery and intelligence, *Mind* **59**: 433–460.

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between men and machines, *Communications of the ACM* **9**: 36–45.

Whitby, B. (1996). The Turing test: AI's biggest blind alley?, *in* P. Millican and A. Clark (eds), *Machines and Thought: The Legacy of Alan Turing*, Vol. 1, Oxford University Press, Oxford, pp. 53–62.

Winch, P. (1980–1981). Eine Einstellung zur Seele, *Proceedings of the Aristotelian Society* **81**: 1–15.

Wittgenstein, L. (1961). *Tractatus Logico-philosophicus*, Routledge & Kegan Paul, London.

Wittgenstein, L. (2009). *Philosophische Untersuchungen/Philosophical Investigations*, 4th edn, Wiley-Blackwell, Oxford.

Zlotowski, J., Proudfoot, D. and Bartneck, C. (2013). More human than human: Does the uncanny curve really matter?, *Proceedings of the HRI2013 Workshop on Design of Humanlikeness in HRI from uncanny valley to minimal design* pp. 7–13.