Where Intelligence Lies: Externalist and Sociolinguistic Perspectives on the Turing Test and AI

Shlomo Danziger(✉)

Department of Cognitive Science, Hebrew University of Jerusalem, Jerusalem, Israel
shlomo.danziger@mail.huji.ac.il

Abstract. Turing's Imitation Game (1950) is usually understood to be a test for machines' intelligence; I offer an alternative interpretation. Turing, I argue, held an externalist-like view of intelligence, according to which an entity's being intelligent is dependent not just on its functions and internal structure, but also on the way it is perceived by society. He conditioned the determination that a machine is intelligent upon two criteria: one technological and one sociolinguistic. The Technological Criterion requires that the machine's structure enables it to imitate the human brain so well that it displays intelligent-like behavior; the Imitation Game tests if this Technological Criterion was fulfilled. The Sociolinguistic Criterion requires that the machine be perceived by society as a potentially intelligent entity. Turing recognized that in his day, this Sociolinguistic Criterion could not be fulfilled due to humans' chauvinistic prejudice towards machines; but he believed that future development of machines displaying intelligent-like behavior would cause this chauvinistic attitude to change. I conclude by discussing some implications Turing's view may have in the fields of AI development and ethics.

Keywords: Alan Turing · Turing Test · Imitation Game · Artificial Intelligence
Externalism

1 Introduction

Can machines be intelligent? In his 1950 paper "Computing Machinery and Intelligence", Alan Turing introduced the Imitation Game (IG) in which a machine tries to imitate human intellectual behavior to such an extent that a human interrogator mistakes the machine for a human. The Imitation Game, later known as the Turing Test, has been commonly understood to be a test for intelligence: A machine that does well in the Game must be regarded as intelligent.

Turing's paper, considered a classic in the fields of AI and philosophy of cognitive science, raises many difficulties, and several attempts have been made throughout the years to explain Turing's intentions (see Saygin et al. 2000; Oppy and Dowe 2011). The commonality between almost all interpretations offered is that they see the IG as a test for *intelligence*. In this essay I reject that widely accepted view and propose an alternative way of understanding Turing's paper and his approach to intelligence. I shall show

V. C. Müller (Ed.): PT-AI 2017, SAPERE 44, pp. 158–174, 2018.
https://doi.org/10.1007/978-3-319-96448-5_15

that Turing holds an externalist-like view of intelligence, which bears resemblance to Wittgenstein's approach to the mental domain (Wittgenstein 2009; 1958). My reading of Turing is based on remarks he makes in other publications (especially Turing 1947; Turing 1948; Turing et al. 1952) and on careful reading of the 1950 paper itself.[1]

In Sect. 2 I introduce the two main streams of interpretation of the IG that have been suggested by Turing's commentators, and I mention some of the problems they raise. In Sect. 3 I discuss some remarks Turing makes in his earlier publications that I believe may shed light on the way he understands the term "intelligence". Sections 4 and 5 are the heart of this essay: In Sect. 4 I present my *technological* interpretation, pointing out what the IG is intended to test and what is outside its scope, and I discuss Diane Proudfoot's interpretation of the IG. In Sect. 5 I present Turing's prediction that in the future the meanings of concepts will change, allowing machines to be deemed "intelligent"; and I offer a critical look at this line of thought. In Sect. 6 I discuss some implications of Turing's approach vis-a-vis AI development and ethics.

2 Imitation Game: Common Interpretations

In his 1950 paper Turing describes the IG as follows: A human interrogator communicates via teletext with another human and with a machine, without knowing which is which. The interrogator must try to ascertain which of the two beings is human by asking each of them any questions whatsoever; each of the beings must try to convince the interrogator that *it* is the human, by answering in a human-like manner. According to the accepted reading, the logical structure of Turing's argument is as follows:

(1) *A machine that does well in the IG – a machine that successfully imitates human intellectual behavior to the extent that the interrogator cannot tell the difference – must be regarded as an intelligent (or a thinking*[2]*) entity*

(2) *Machines that do well in the IG can indeed be constructed*

Therefore,

(3) *Intelligent machines are possible*

The role of the IG within the argument seems puzzling, and several attempts have been made to explain Turing's paper. Following Diane Proudfoot's classification (2013),

[1] My reading is supported by several pieces of non-orthodox commentaries of Turing scattered throughout the literature, such as Whitby (1996), Boden (2006, pp. 1346–1356), Sloman (2013), and especially Proudfoot (2005; 2013).
Some of the arguments suggested in this paper have appeared in Danziger (2016).

[2] As Piccinini (2000), Proudfoot (2013), and others have pointed out, Turing uses the terms "thought" and "intelligence" interchangeably. Although I will not differentiate between the terms, for reasons of uniformity I shall usually use the term "intelligence".

I shall point out two main streams of interpretation suggested in the literature and briefly mention some of the problems they raise.[3]

2.1 Behavioristic Interpretations

According to *behavioristic* interpretations, Turing held that “intelligent-like behavior” is the *definition* of intelligence: Any system (that is, any organism or machine) whose behavior is similar to that of an intelligent entity (a human) is itself intelligent. French (1990, p. 53) exemplifies this operational definition of intelligence by saying that according to Turing, “[w]hatever *acts* sufficiently intelligent *is* intelligent.”[4]

Behavioristic interpretations – the most common way of understanding Turing’s paper – raise several difficulties. For according to these interpretations, Turing seems to be going against a very basic human intuition – that mental occurrences and properties are *internal* traits, independent of *external* actions.[5] Also, it is not clear why, according to Turing, intelligence is tested for by verbal behavior, and not by any other human cognitive faculty.

2.2 Inductive Interpretations

As opposed to behavioristic interpretations, *inductive* interpretations claim that Turing indeed sees intelligence as an *internal* property of a system – one that takes place inside it, and not as an external, behavioristic trait. According to these interpretations, Turing holds that a system’s success in the IG gives us *good grounds* to assume it possesses the property of intelligence, based on the success that this kind of attribution has shown hitherto; and in the absence of contradicting evidence, we should regard such a system as intelligent.[6]

Inductive interpretations raise problems too, for according to them Turing implies the following: “Due to our long-learned experience, we humans *tend* to attribute (the internal property of) intelligence to systems on the basis of their (external) behavior; therefore, we *must* attribute intelligence to systems that display intelligent behavior (i.e.,

[3] Almost all commentaries on – and attacks against – Turing’s paper can be classified into one of the two streams of interpretation described in Sects. 2.1 and 2.2; see Proudfoot (2013) for detailed analysis and critique of these interpretations. Other ways of classification can be found in Saygin et al. (2000) and in Oppy and Dowe (2011).

[4] Also Searle’s interpretation of Turing is behavioristic: “The Turing Test is typical of the tradition in being unashamedly behavioristic and operationalistic” (Searle 1980, p. 423; cf. next footnote). References to other behavioristic interpretations can be found in Proudfoot (2013), Copeland (2004, pp. 434–435), and Moor (2001, pp. 81–82).

[5] This is the crux of perhaps the two most well-known arguments against the IG, namely, Searle’s Chinese Room (Searle 1980) and Block’s Blockhead / Aunt Bubbles Machine (Block 1981; 1995): Intelligence, they maintain, cannot be captured in behavioral terms alone. (Note that both arguments belong to the behavioristic school of interpretation, in that they assume that the IG is intended to be a behavioral test for intelligence.)

[6] The main proponent of the school of inductive interpretations is Moor (1976; 2001). Other inductive interpretations can be found in Watt (1996) and Schweizer (1998).

do well in the Game).” This move from “we tend” to “we must” seems strange, for a system’s external behavior can be misleading, causing us – the observers – to adopt a false picture of the system’s inner state; why *must* we trust behavior so blindly?

3 Turing 1947 and 1948: Machines Can Think

Before I present my interpretation I would like to refer to Turing’s earlier publications from 1947 and 1948, which can shed light on his approach and give us a better understanding of his 1950 paper.

3.1 Intelligence as an Emotional Concept[7]

Towards the end of his 1948 paper, after a lengthy discussion of ways in which machinery can imitate various human cognitive functions, Turing writes (1948, p. 431, my italics):

> **Intelligence as an emotional concept**
>
> The extent to which we regard something [a machine or an organism, SD] as behaving in an intelligent manner is determined as much by *our own state of mind and training* as by the properties of the object under consideration. If we are able to explain and predict its behaviour or if there seems to be little underlying plan, we have little temptation to imagine intelligence. With the same object therefore it is possible that one man would consider it as intelligent and another would not; the second man would have found out the rules of its behaviour.

Turing could have said that in a case of different epistemic viewpoints one of the viewers would be *wrong*, but he chose to say otherwise: A given system could be both “intelligent” and “non-intelligent” at the same time if viewers having two such opposing viewpoints existed. Contrary to the way he was understood by behavioristic and inductive interpretations, Turing implies here that “intelligence” cannot be given a clear-cut definition in terms of *behavior* or in terms of *internal properties*; one system having a single set of behavior and internal properties allows for two opposing viewpoints, and neither would be wrong.

In saying that intelligence is an “emotional concept” Turing is referring to the emotions and reactions of the people perceiving the system, and thus points out the major role of the environment in deeming a system “intelligent”. Intelligence, so to speak, is in the eye of the beholder: What defines certain systems as intelligent is, first and foremost, the fact that we humans perceive *those* systems, and not others, as intelligent. All systems – including intelligent ones – are mere physical mechanisms; what makes a system unique and defines it as “intelligent”, says Turing, is the viewpoint of the people in its environment.

[7] The ideas in this section draw partly on Proudfoot (2005; 2013). As I shall show later, my inter pretation of Turing differs from Proudfoot’s in small but crucial points; to prevent inaccuracies I shall refrain for now from mentioning her take on the subjects discussed, despite my great debt to her work.

Note that Turing's methodology reveals his approach to intelligence. Turing adopts a Wittgensteinian-like methodology in his analysis of the term "intelligence", reviewing cases in which people would or would not perceive systems as intelligent and attribute intelligence to them (Turing 1947, p. 393; 1948, pp. 412, 431). But by limiting the discussion to analysis of humans' *reactions* to systems and to the way the term "intelligence" is used in ordinary language, and by refusing to provide any further definition of intelligence (cf. Turing et al. 1952, p. 494), Turing reveals that in his view, what matters is the question of whether machines would be perceived as intelligent by human society. If they would be – then they could be said to be "intelligent", and no further inquiry would be needed (i.e., there would be no need to ask if they are "really" intelligent, according to some real-but-unknown definition that exists "out there").

Turing's approach thus described highly resembles what Coeckelbergh (2010) and Torrance (2014) call the *social-relationist* perspective (as opposed to the *realist* perspective).[8] Following this terminology, Turing's approach can be formalized as the following *premise*:

Social-Relationist Premise: *A system (organism or machine) perceived as intelligent by human society is an intelligent system*[9]

Now, given that "an intelligent system" is logically equivalent to "a system that is *perceived* as intelligent" (*Social-Relationist Premise*), the question

(Q1) *Can machines be intelligent?*

can be rephrased as

(Q2) *Is it possible for machines to be perceived as intelligent?*

In order to answer question (Q2) we must first find what *causes people* to perceive certain systems as intelligent:

(Q2.1) *What would be a sufficient condition for a system to be perceived as intelligent?*[10]

Hence, in our attempt to understand the criteria for intelligence, we find that before delving into questions in the domains of cognition and computation pertaining to the system's structure and functions, we must first focus on the fields of sociology and psychology, and ask questions pertaining to the people in the system's environment: What causes *human society* to regard a system – organism or machine – as intelligent? Once we answer the sociological questions we can proceed to the technological ones, regarding the system's functions and structure. Let us ask, therefore: What would be a sufficient condition for a system to be perceived as intelligent? What properties or abilities would an intelligent machine have?

[8] Turing's approach bears resemblance also to Dennett's "intentional stance" (Dennett 1987a).

[9] In Sects. 3.2 and 4.3 I shall bring further textual evidence for this being Turing's approach, and shall briefly discuss what might have motivated Turing into adopting such a stance.

[10] Turing is trying to *prove* that the existence of an intelligent machine is possible, and is not merely *asking* if it possible. Therefore he will try to show that machines fulfill a *sufficient* condition for being (perceived as) intelligent, and will put less emphasis on the *necessary* conditions.

3.2 Sufficient Conditions for Intelligence

In the passage titled "Intelligence as an emotional concept" quoted above (Sect. 3.1), Turing claims that when we encounter a simple system, one that "we are able to explain and predict its behaviour," we have "little temptation to imagine intelligence," and so we experience it as a mere mechanistic, non-intelligent system (Turing 1948, p. 431). According to Turing, the reason no machine has ever been perceived by humans as intelligent is that all machines that humans have ever encountered were of very limited character (1948, p. 410); no machine had ever displayed sophisticated human-like cognitive abilities involving learning, such as the ability to learn from experience and the ability to modify one's own "programming" (1947, pp. 392–393). But if, says Turing, such a "learning machine" were built – it *would* be experienced by humans as intelligent (1947, p. 393, my italics):

> Let us suppose we have set up a machine with certain initial instruction tables [programs, SD], so constructed that these tables might on occasion, if good reason arose, modify those tables. One can imagine that after the machine had been operating for some time, the instructions would have altered out of all recognition, but nevertheless still be such that one would have to admit that the machine was still doing very worthwhile calculations. Possibly it might still be getting results of the type desired when the machine was first set up, but in a much more efficient manner. In such a case one would have to admit that the progress of the machine *had not been foreseen* when its original instructions were put in. It would be like a pupil who had learnt much from his master, but had added much more by his own work. When this happens *I feel that one is obliged to regard the machine as showing intelligence.*

Note the last sentence: According to Turing, a "learning machine", programmed in such a sophisticated way that it could modify its own code, would arouse a feeling of surprise in its observers, and they would find themselves *regarding it* as showing intelligence. In his 1948 paper Turing repeats this prediction when discussing the would-be reaction of a human playing chess against a machine (as part of an early version of the IG; p. 431). In fact, he seems to say that he himself had actually reacted in such a way when encountering a chess-playing machine (1948, p. 412). Turing's descriptions of these would-be reactions (or actual reactions) of humans to such machines amount to his claiming that the conjunction of the properties of a "learning machine" is a *sufficient condition* for this machine to be perceived as an intelligent system:[11]

Sociological Claim: *A "learning machine" would be perceived by human society as intelligent*

Another important claim Turing makes in his 1947 and 1948 publications is that "learning machines" like the one just discussed *can* be built. According to Turing, the digital computer – which was then being developed – could carry out any task that the human brain could, and could therefore display the human-like cognitive abilities needed for being a "learning machine". (The digital computer will be discussed later in greater length.) As opposed to the "very limited character of the machinery which has been used

[11] There is no need to point out here which properties of the "learning machine" are necessary conditions for perceiving a system as intelligent; all that is being claimed is that a "learning machine" indeed *has* these properties, whatever they may be.

until recent times" which "encouraged the belief that machinery was necessarily limited to extremely straightforward, possibly even to repetitive, jobs" (Turing 1948, p. 410), the digital computer *would* be able to learn from experience, change its own programming, and display any other property of a "learning machine":

Minor-Technological Claim: *It is possible to build a "learning machine"*[12]

3.3 The Logical Structure of Turing's Argument

The logical structure of Turing's argument in his 1947 and 1948 papers is as follows:

Minor-Technological Claim: *It is possible to build a "learning machine"*

Sociological Claim: *A "learning machine" would be perceived by human society as intelligent*

– ***Conclusion:*** *It is possible to build a machine that would be perceived by human society as intelligent*

Social-Relationist Premise: *A system (organism or machine) perceived as intelligent by human society is an intelligent system*

– ***Conclusion:*** *It is possible to build an intelligent machine* ***(Q.E.D)***

To sum up: In his 1947 and 1948 publications, Turing argues that there *can* be intelligent machines. He claims that the construction of "learning machines" is a technological challenge that *can* be met (*Minor-Technological Claim*), and claims that these machines would inevitably be *perceived* as intelligent by human society (*Sociological Claim*) and would thereby *be* "intelligent machines" (*Social-Relationist Premise*). Turing shifts the focus from technological questions regarding the system's internal structure to sociological questions regarding the people in the system's environment. In doing so, he sidesteps the need to define "intelligence" (or "thought"); regardless of what the definition of intelligence is, if human society were to perceive a machine as intelligent – it would be correct to say that it *is* intelligent.

4 Turing 1950: Technological Interpretation

I shall now turn to analyze Turing's famous 1950 paper, where he introduces the well-known Imitation Game. I shall offer my "technological" interpretation and claim that in 1950 Turing retreats from the stance he presented in 1947 and 1948, realizing that the *Sociological Claim* (Sect. 3.2), according to which machines with special functions would be perceived by society as intelligent entities, was naïve and perhaps too optimistic.

[12] It will later become clear why this claim is labeled "minor".

4.1 From Specific Abilities to All-Encompassing Imitation Ability

In the 1947 and 1948 publications discussed above, Turing claims that if a machine were to display the abilities of a "learning machine" (learning from experience, reprogramming itself, etc.), it would inevitably be perceived as intelligent by human society (*Sociological Claim*); and he claims that machines *can*, in principle, display these abilities (*Minor-Technological Claim*). In his 1950 paper, though, Turing aims higher. He no longer tries to convince the reader that the digital computer could imitate *some ability or another* (however important that ability may be for being considered "intelligent"), but claims that the digital computer can imitate the entire human cognitive system, as it can imitate the human brain *as a whole.*[13]

Turing's confidence that machines could do so is based on the strong imitation ability of the "universal machine" (later known as the "Turing machine") introduced in his 1936 paper. According to Turing, each function of the human brain could be imitated *closely enough* by a "digital state machine"; all functions of all digital state machines could be fully imitated by a universal machine; hence, all functions of the human brain could be imitated closely enough by a universal machine. This machine could, in principle, do anything a human brain can do; it could successfully carry out any human cognitive task.

Turing thus moves from discussing machines that could display specific, unique abilities that would be sufficient for intelligence ascription (1947, 1948), to discussing machines that could imitate the entire human cognitive system (1950). Machines of the latter kind would have all the abilities that machines of the former kind had, and many more. If such brain-imitating machines (machines of the latter kind) were built, we could answer question (Q2) above – "Is it possible for machines to be perceived as intelligent?" – with an unequivocal "yes", without needing to determine which abilities, exactly, are "responsible" for a system's being perceived as intelligent. For *whatever* these abilities might be – we would know for certain that they *could* be realized by these machines that can do everything the human brain does.

4.2 The Imitation Game and the Technological Criterion

From a technological aspect, these machines that satisfactorily imitate any brain function would be quite sophisticated. The Imitation Game is a means to check if the *technological challenge* of building such sophisticated machines has been met. The Game tests if a given machine acts so much like a human brain that one cannot differentiate between the two; in other words, it tests if a machine's behavior is *intelligent-like*.[14]

A machine that does well in the IG – a machine that has intelligent-like behavior – fulfills what I call the *Technological Criterion* for intelligence. The Technological Criterion requires that a system's structure (or program) enables it to imitate the human brain very well, to the extent that a human interrogator experiences it as

[13] Hodges (2014, p. 530) explains in a similar way the difference between Turing's 1948 and 1950 papers.

[14] "Intelligent-like behavior" may be roughly defined as "behavior that under regular circumstances cannot be differentiated from that of a human".

intelligent. The IG, therefore, tests if a given system fulfills the Technological Criterion for intelligence.[15]

Turing designed the IG as a test involving verbal interaction because of the huge technological challenge this posed. Constructing machines that successfully engage in real-time human conversation seemed to him as an extremely difficult task from a technological/algorithmic aspect, on par with – if not harder than – constructing machines that possess any other cognitive ability related to “learning from experience” (Sect. 3.2). Turing, I claim, saw the IG as an “AI-complete” problem (Mallery 1988, p. 47, fn. 96): If a machine could do well in the Game, it could accomplish practically *anything* related to AI. (Even today, programming a computer to do well in Natural Language Processing tasks is considered one of the greatest challenges in AI development.) According to Turing, if a machine were constructed in such a way that it displayed intelligent-like behavior and caused a human interrogator to experience it as intelligent, its engineers could lean back in satisfaction knowing that the technological challenge of creating brain-imitating, intelligent-like machines had been met. It is with regard to this *technological* challenge that Turing makes the following prediction (1950, p. 442):

> I believe that in about fifty years’ time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70% chance of making the right identification after five minutes of questioning.

In 1950, then, Turing predicts that brain-imitating machines with *intelligent-like* behavior could indeed be built. Did he claim that these machines would be *intelligent*?

4.3 The Iron Curtain of Sociolinguistic Restrictions and the Sociolinguistic Criterion

Turing’s 1947 and 1948 publications imply that “learning machines” that displayed abilities such as learning from experience, reprogramming themselves, etc. would be perceived by human society as intelligent (*Sociological Claim*, Sect. 3.2) and would thus *be* intelligent (*Social-Relationist Premise*, Sect. 3.1). One would therefore expect that in his 1950 paper Turing would say the same of machines that fulfilled the Technological Criterion and displayed intelligent-like behavior; those machines, one would presume, would surely be perceived by society as intelligent entities. But careful reading reveals that in his 1950 paper Turing retreats from the naïve view presented in his earlier publications. He realizes that while humans perceive other humans as “intelligent entities”, humans perceive machines *a–priori* as “non-intelligent entities”, due to a *chauvinistic attitude* towards machines that humans have (and may be unaware of). Turing recognizes that even if a machine were to fulfill the Technological Criterion for intelligence by doing well in the IG, *human prejudice* would preclude any possibility of machines being thought of as intelligent. In a 1952 radio broadcast Turing describes this prejudiced attitude (Turing et al. 1952, p. 500, my italics):

[15] The *Technological Criterion* (1950) is closely connected to the *Minor-Technological Claim* (1947, 1948) but is more “demanding” (as explained above, Sect. 4.1); that is why the 1947–1948 claim is labeled “minor”.

> If I had given a longer explanation [of how to construct a machine that would have the ability to identify analogies, SD] … you'd probably exclaim impatiently, 'Well, yes, I see that a machine could do all that, but *I wouldn't call it thinking*.' 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.*

Turing realizes that the term "intelligent machine" is an *oxymoron*: People think of machines as systems whose workings they can understand; and a system whose workings could be understood is seen as consisting of *mere mechanistic processes*, devoid of any intelligence.[16] Turing hence recognizes that the concept of intelligence could not be ascribed to machines, by definition. This idea strongly resembles one that appears in Wittgenstein's later writings (2009: §360):

> But surely a machine cannot think! – Is that an empirical statement? No. We say only of a human being and what is like one that it thinks. We also say it of dolls; and perhaps even of ghosts.[17]

It seems that according to Wittgenstein, even if there were a person who did not have that a-priori chauvinistic attitude towards machines and *did* see them as potentially intelligent systems (as Turing might have), that person would run into the iron curtain of language conventions that prevent us from applying the term "intelligent" to machines in the literal sense. Machines cannot be said to be intelligent, Wittgenstein would say, because of the way the terms "intelligent" and "machine" are used in language. This, presumably, is the reason why Turing, despite his being convinced that machines could do *anything* a brain could and could behave in an intelligent-like manner, refrains in his 1950 paper from explicitly stating that such machines would be *intelligent*.

To frame it differently: Turing understood that alongside the Technological Criterion for intelligence, there also exists what I call the *Sociolinguistic Criterion*: the requirement that the system be such that its kind is perceived by society as potentially intelligent. According to the Sociolinguistic Criterion, a system that is perceived a-priori by society as non-intelligent (i.e., belongs to a species or a kind that is perceived a-priori as non-intelligent) cannot be said to be intelligent, by definition. Turing realized that in the year 1950, the Sociolinguistic Criterion, which is dependent on the system's environment (human society), could not be fulfilled with regard to machines. Doing well in the IG – fulfillment of the Technological Criterion – would show only that the system's behavior is *intelligent-like*, but this would not break the sociolinguistic barricade seeded in the minds of humans that causes them to see machines a-priori as non-intelligent entities.[18]

[16] Bringsjord et al. (2001) mention a similar idea of "restricted epistemic relation": They suggest the "Lovelace Test" for intelligence in which "not knowing how a system works" is a necessary condition for attributing intelligence to it. The fundamental difference between the Lovelace Test and Turing's IG will be explained later (fn. 23).

[17] Other clear remarks of Wittgenstein in this spirit are Wittgenstein (2009, §281) and Wittgenstein (1958, p. 47). The similarity between Turing's and Wittgenstein's ideas here has been pointed out also by Boden (2006, p. 1351) and Chomsky (2008, p. 104).

[18] The *Sociolinguistic Criterion* (1950) is closely connected to the *Sociological Claim* (1947, 1948) mentioned in Sect. 3.2. The addition of the "linguistic" component will soon be explained.

This is how we should understand Turing's enigmatic remark (pun intended) in his 1950 paper, which follows his prediction quoted above in Sect. 4.2 (Turing 1950, p. 442, my italics):

> The original question, "Can machines think?" I believe to be *too meaningless* to deserve discussion.

This question, says Turing, is "meaningless" – in the Wittgensteinian sense: Machines are not things that fall under the concept of "thinking" (or "intelligence"). And this is why Turing, at the outset of his paper, replaces the question "Can machines think?" with the question of whether or not machines can do well in the IG. While Turing's original question touches on both Technological and Sociolinguistic Criteria (and contains a built-in negative answer), the new question relates to the Technological Criterion alone; the IG offers a way to check if the Technological Criterion has been satisfied.[19]

To recap: Turing tackles the question "Can machines think?" by saying, "Look, a machine can do anything a brain can do. Anything. It can behave so similar to a human brain that it can even do well in the Imitation Game. Does this mean that a machine can *think*? No. But that is not because there is something it cannot *do*; it's not like the fact that I can't climb a very steep cliff due to the limits of my strength. A machine cannot think because the term 'thinking machine' is an *oxymoron*; it cannot be said to think because of the way the terms 'think' and 'machine' are used in language."

4.4 Proudfoot's Interpretation: The Imitation Game as a Test for Intelligence

Before presenting the last stage in Turing's argument I must mention the writings of Diane Proudfoot (2005; 2013), which greatly inspired my interpretation presented thus far. In her comprehensive and enlightening papers, Proudfoot promotes an externalist-like interpretation of Turing, according to which intelligence is (what she calls) a *response-dependent* property (Proudfoot 2013, p. 398):

> Turing's remarks suggest something like this schema: *x* is *intelligent* (or *thinks*) if, in an unrestricted computer-imitates-human game, *x* appears intelligent to an average interrogator.

[19] At this point one might raise the following objection: "Your reading boldly ignores the next sentence in Turing's paper, in which he supposedly predicts that in fifty years there would be intelligent machines (1950, p. 442): 'Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.' This implies that Turing identified success in constructing machines that do well in the Game – with success in creating *intelligent* machines; the timeframe in both sentences is the same (the year 2000), and so they seem to be referring to the *same* futuristic occurrence!" My reply, in short, is that this objection is based on an incorrect – albeit very common – reading of the passage in Turing's paper. Turing, I claim, makes two different predictions here, and these predictions are connected *causally* but not *logically*. "Doing well in the IG" is not the same as "being intelligent". The IG, I insist, is not a test for intelligence, but a test only for the Technological Criterion of intelligence: it tests if a system's behavior is *intelligent-like*. (I shall return to this issue in Sect. 5.1.)

Turing, according to this, holds that what defines a system as intelligent is the attitude of an *average interrogator*, who is supposed to represent society in an unadulterated, impartial way (like a jury in court, perhaps).

My interpretation is close to Proudfoot's but differs from it in crucial points. The main difference is that her interpretation, like behavioristic and inductive ones mentioned earlier, sees the IG as a test for *intelligence*. But Turing, I claim, did not intend the Game to be a test for intelligence. Intelligence requires that *society* perceive the system as intelligent, and the IG does not test *that*. It tests only whether a *single, isolated* interrogator *temporarily* experiences the system as intelligent during the few moments in which the interrogator does not yet know that s/he is conversing with a machine. Turing himself refers to the IG as an "imitation test" (Turing et al. 1952, p. 503; cf. p. 495); indeed, the Game is a test for the Technological Criterion only, a test for *intelligent-like behavior*.[20] A test for *intelligence*, on the other hand, would require the fulfillment of the Sociolinguistic Criterion too.[21]

5 Shifts in the Meanings of Concepts

5.1 Turing's Prediction

According to my reading, in 1950 Turing acknowledged that machines could not "be intelligent" or "think", due to humans' prejudiced attitude towards machines and the way the terms "intelligence", "thinking" and "machine" were used in language. But the way people use words can change. Here is Turing's remark quoted above (Sect. 4.3) followed by his prediction (1950, p. 442, my italics):

> The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century *the use of words and general educated opinion will have altered so much* that one will be able to speak of machines thinking without expecting to be contradicted.

Turing believed that technological progress – development of machines that do well in the IG and show intelligent-like behavior – would eventually cause humans' chauvinistic attitude towards machines to erode. The term "intelligent machine" would then no longer constitute an oxymoron, as the meanings of the concepts "intelligence" and

[20] Aaron Sloman, too, sees the IG as Turing's way of defining a *technological* challenge, and not as a test for intelligence (Sloman 2013). In an earlier version of his paper Sloman expresses his dissatisfaction with the orthodox interpretations of the IG; I found myself wholly identifying with his words (my italics): "It is widely believed that Turing proposed a test for intelligence. This is false. *He was far too intelligent to do any such thing,* as should be clear to anyone who has read his paper…"
(Source: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/turing-test.html. Accessed Oct. 11, 2017.)

[21] To develop this point further: A real test for a system's intelligence would check if the system is perceived as intelligent by society as a whole, in an ongoing manner, in normal life situations. But if that were to happen there would be no need for an intelligence test, because "society perceiving a system as intelligent" is the *definition* of a system's being intelligent, not a *sign* of it! (See the *Social-Relationist Premise*, Sect. 3.1.)

"machine" would have changed; the concept of "intelligence" would then be applicable to machines.[22] In that future state, the Sociolinguistic Criterion would have been fulfilled, as society would indeed see machines as *potentially intelligent* systems. If a certain machine also fulfilled the Technological Criterion, it would be perceived as an intelligent system, and would rightly be said to be an intelligent machine.

In conclusion, Turing thought both that machines could do well in the IG *and* that intelligent machines were possible. But contrary to the accepted reading presented in Sect. 2, Turing saw the connection between the IG and intelligence not as a *logical* connection, but as a *causal* one. He did not claim that machines that do well in the IG *are* intelligent, but that success of machines in the IG would eventually *cause* people to see machines as intelligent.

The widespread misunderstanding of the IG can be further clarified by differentiating between *descriptive* and *normative* readings. While my interpretation sees Turing's account as descriptive ("That is how people *would* react upon their encounter with machines that do well in the IG"), Turing's commentators – who thought he intended the IG to be a test for intelligence – understand him as giving a normative account ("That is how we *should* regard machines that do well in the IG"). I am of the opinion that the normative reading is an incorrect understanding of Turing's paper.[23]

5.2 A Critical Look at Turing's Prediction

Technically speaking, Turing was too optimistic; the year 2000 has passed and we still do not perceive of machines as thinking/intelligent entities. (In fact, it has been stressed that the only time we say of a computer that it is "thinking" is when it gets stuck.) Turing's prediction that the meanings of concepts will change may indeed come about sometime in the future. However, I want to suggest the opposite scenario: If we develop machines that have intelligent-like abilities and act very much like humans, we might stop identifying those abilities with intelligence, just like we stopped seeing "winning the chess game" as a sign for intelligence in 1997, when "Deep Blue" beat chess champion Kasparov.[24] "Tesler's Theorem" expresses this point elegantly (Larry Tesler, ca. 1970):

> Intelligence is whatever machines haven't done yet.

[22] This is how Turing's prediction was understood by Mays (1952, pp. 149–151), Beran (2014) and others. (Piccinini 2000 understands that Turing *hopes* such a change will occur.) For an illuminating discussion regarding the possibility of this sort of change (not concerning Turing's paper) see Torrance (2014).

[23] The main difference between Turing's IG and Bringsjord et al.'s "Lovelace Test" mentioned above (fn. 16) is that while the IG is descriptive, the Lovelace Test is normative (see Bringsjord et al. 2001, p. 9).

[24] Sloman makes a similar point and says that while computers are now doing much cleverer things, "increasing numbers of humans have been learning about what computers can and cannot do" (Sloman 2013, p. 3). Indeed, getting humans to attribute intelligence to machines might become harder with time.

Hence, an engineer might do everything philosophers said should be done in order to develop intelligent systems – only to discover that the philosophers keep changing the rules.

Moreover: If machines start acting like humans, we humans might find ourselves changing the way *we* behave in order to distance ourselves from machines so that we remain the "superior race". Our new behavior will then become the new standard for intelligence, the behavior in virtue of which we perceive systems as intelligent. In such a scenario, humans would make sure to act in a unique way so that society (whomever that may include) clearly understands that humans, and not machines, are the *real* bearers of intelligence.

6 Implications of Turing's View

6.1 Externalism and AI Development

When discussing the criteria for intelligence, Turing focuses on the way a system is perceived by human society, rather than on the system's functions or internal structure. Turing, therefore, can be said to hold an *externalist-like* view of intelligence (and of the mental domain in general[25]). A system's functions and internal structure may indeed play an important role in shaping society's attitude towards the system (thereby circuitously contributing to the definition of the system as "intelligent"), but by no means are they the only factors.

I think Turing's approach may lead to interesting insights regarding the ongoing attempt to develop intelligent systems. Recent years have seen efforts in the fields of technology and algorithm development to devise human-like intelligent systems (including ongoing attempts to write computer programs that would "pass the Turing Test"). Turing's approach teaches us that it would be wise to pay attention also to the major role that *society* plays in determining the intelligence of a system. This might lead developers to put more emphasis on properties that had once been considered irrelevant to intelligence. One such property is the external appearance of the system. Another is the way the system was developed: Humans might be more inclined to attribute intelligence to a system that, like themselves, went through a long and tedious learning process, as opposed to a system that had a whole database injected into it; the latter might seem less human-like and would be less likely to be perceived as intelligent.[26]

Awareness of the sociological dynamics involved in determining a system's intelligence may also teach us why we must have *patience* when trying to construct intelligent

[25] In his brief reply to the "Argument from Consciousness", Turing seems to claim that if a machine did well in the IG it would be *perceived* as conscious too (1950, pp. 445–447; see Michie 1993, pp. 4–7. But cf. Copeland 2004, pp. 566–567). I am of the opinion that likewise intelligence, also consciousness and other mental phenomena can be explained in terms of being perceived by society; I plan to discuss this elsewhere.

[26] Both properties mentioned were suggested by Mays (1952), in his analysis of Turing's 1950 paper. Interestingly, Turing himself seems to have viewed both properties as insignificant for intelligence attribution (see Turing 1950, p. 434; Davidson 1990). For a list of other properties that might shape humans' attitude towards machines, see Torrance (2014).

systems. As pointed out by Beran (2014, pp. 54–55), for machines to be intelligent, humans must first adapt to the idea of intelligent machinery, and this change of attitude may take time. In addition, developers should be willing to accept that humans' *stubborn chauvinistic attitude* might completely prevent the possibility of perceiving machines as intelligent. If machines were to acquire abilities considered paradigmatic intelligent-like behavior (such as learning from experience), people might stop seeing those abilities as central to intelligence, and "replace" them with others. This would be equivalent to changing the criteria for intelligence, rendering machines as *non-intelligent* again and again (*a-la* Tesler's Theorem), every time it seems as though they "almost got there."

6.2 Ethics

According to Turing, if humans' chauvinistic attitude towards machines changed and they came to see some machines as intelligent – those machines would *really be* intelligent. But what if only part of society came to see machines as intelligent beings (or, for the sake of the argument, as conscious beings), while the other part kept seeing them as mere machinery? According to Turing's approach, these two points of view would reflect two incommensurable paradigms (to use Thomas Kuhn's terminology), and there would not be any objective viewpoint from which this dispute could be settled. Human society would then be split over the question of how human-like machines should be treated; for example, should they be given human(!) rights and be freed from slavery? This question would probably not be resolved by logical reasoning, but by persuasion, or perhaps by violence (among humans). Indeed, due to the ethical aspects involved, people would probably have very little tolerance for the "other" opinion, which they would see as a totally unethical stance. In addition, the ethical flavor of the dispute would not leave much room for personal ambivalence, as each person would feel that they *must* take a side in the debate.[27]

7 Epilogue

Turing illuminates the important role played by human society in determining whether machines are intelligent. Machines cannot be perceived as intelligent in a society that has a prejudiced chauvinistic attitude towards them; but if this a-priori attitude were to change, brain-imitating machines could indeed be perceived as intelligent entities. Turing, who was convinced that machines could do everything a human brain does, feared that *his* opinion would not be accepted due to human prejudice towards himself, as appears in a worried letter he wrote in 1952 while standing trial on charges of "gross indecency" (Hodges 2014, pp. xxix–xxx):

[27] Discussions regarding the active role of humans in drawing the borders of the "Charmed Circle" of consciousness or intelligence (relevant also to the issue of animal consciousness and to disputes regarding humans' attitude towards animals) can be found in Dennett (1987b) and Michie (1993).

> I'm rather afraid that the following syllogism may be used by some in the future –
> Turing believes machines think
> Turing lies with men
> Therefore machines cannot think

When studying the nature of thought and intelligence, concentrating solely on the system's functions and internal structure can be misleading. That is not where intelligence lies. In emphasizing the major role of society, Turing's research – while focusing on machine intelligence – can teach us quite a bit about human intelligence as well.

Acknowledgments. Research for this paper was financially supported by the Sidney M. Edelstein Center for the History and Philosophy of Science, Technology, and Medicine at the Hebrew University of Jerusalem; and by the Centre for Moral and Political Philosophy (CMPP) at the Hebrew University of Jerusalem. I thank Orly Shenker, Oron Shagrir, Netanel Kupfer, Sander Beckers, Anna Strasser, Selmer Bringsjord and Vincent C. Müller, who reviewed this paper and added thoughtful comments. Special thanks to the participants of the PT-AI 2017 Conference for thought-provoking discussions, and to Shira Kramer-Danziger for her assistance in editing and her wise advice.

References

Beran, O.: Wittgensteinian perspectives on the Turing test. Studia Philosophica Estonica **7**(1), 35–57 (2014)

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

Block, N.: The mind as the software of the brain. In: Smith, E.E., Osherson, D.N. (eds.) Thinking, pp. 377–425. MIT Press, Cambridge (1995)

Boden, M.A.: Mind as Machine: A History of Cognitive Science. Oxford University Press, Oxford (2006)

Bringsjord, S., Bello, P., Ferrucci, D.: Creativity, the Turing test, and the (better) Lovelace test. Mind. Mach. **11**, 3–27 (2001)

Chomsky, N.: Turing on the "imitation game". In: Epstein, R., Roberts, G., Beber, G. (eds.) Parsing the Turing Test, pp. 103–106. Springer, New York (2008)

Coeckelbergh, M.: Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf. Technol. **12**(3), 209–221 (2010)

Copeland, B.J. (ed.): The Essential Turing. Oxford University Press, Oxford (2004)

Danziger, S.: Can computers be thought of as thinkers? Externalist and linguistic perspectives on the Turing test. MA thesis, Hebrew University of Jerusalem (2016). [Hebrew]

Davidson, D.: Turing's test. In: Said, K., Newton-Smith, W., Viale, R. Wilkes, K. (eds.) Modelling the Mind, pp. 1–12. Clarendon Press, Oxford (1990)

Dennett, D.C.: The Intentional Stance. MIT Press, Cambridge (1987a)

Dennett, D.C.: Consciousness. In: Gregory, R.L., Zangwill, O.L. (eds.) The Oxford Companion to the Mind, pp. 160–164. Oxford University Press, Oxford (1987b)

French, R.M.: Subcognition and the limits of the Turing test. Mind **99**(393), 53–65 (1990)

Hodges, A.: Alan Turing: The Enigma. Princeton University Press, Princeton and Oxford (2014)

Mallery, J.C.: Thinking about foreign policy: finding an appropriate role for artificially intelligent computers. The 1988 Annual Meeting of the International Studies Association, St. Louis (1988). 10.1.1.50.3333

Mays, W.: Can machines think? Philosophy **27**, 148–162 (1952)

Michie, D.: Turing's test and conscious thought. Artif. Intell. **60**(10), 1–22 (1993)

Moor, J.H.: An analysis of the Turing test. Philos. Stud. **30**, 249–257 (1976)

Moor, J.H.: The status and future of the Turing test. Mind. Mach. **11**, 77–93 (2001)

Oppy, G., Dowe, D.: The Turing test. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2011). plato.stanford.edu/archives/spr2011/entries/turing-test. Accessed 13 Oct 2017

Piccinini, G.: Turing's rules for the imitation game. Mind. Mach. **10**, 573–582 (2000)

Proudfoot, D.: A new interpretation of the Turing test. Rutherford J. N. Z. J. Hist. Philos. Sci. Technol. 1 (2005). Article 010113. rutherfordjournal.org/article010113.html. Accessed 1 Nov 2017

Proudfoot, D.: Rethinking Turing's test. J. Philos. **110**(7), 391–411 (2013)

Saygin, A., Cicekli, I., Akman, V.: Turing test: 50 years later. Minds Mach. **10**, 463–518 (2000)

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

Searle, J.R.: Minds, brains, and programs. Behav. Brain Sci. **3**, 417–424 (1980)

Sloman, A.: The mythical Turing test. In: Cooper, S.B., Van Leeuwen, J. (eds.) Alan Turing: His Work and Impact, pp. 606–611. Elsevier, Amsterdam (2013)

Torrance, S.: Artificial consciousness and artificial ethics: between realism and social relationism. Philos. Technol. **27**(1), 9–29 (2014)

Turing, A.M.: On computable numbers, with an application to the Entscheidungsproblem. Reprinted in Copeland (2004), pp. 58–90 (1936)

Turing, A.M.: Lecture on the automatic computing engine. Reprinted in Copeland (2004), pp. 378–394 (1947)

Turing, A.M.: Intelligent machinery. Reprinted in Copeland (2004), pp. 410–432 (1948)

Turing, A.M.: Computing machinery and intelligence. Mind **50**, 433–460 (1950)

Turing, A.M., Braithwaite, R., Jefferson, G., Newman, M.: Can automatic calculating machines be said to think? Reprinted in Copeland (2004), pp. 494–506 (1952)

Watt, S.: Naive psychology and the inverted Turing test. Psycoloquy **7**(14) (1996). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.43.2705&rep=rep1&type=pdf. Accessed 2 Nov 2017

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

Wittgenstein, L.: The Blue and Brown Books: Preliminary Studies for the "Philosophical Investigations". Harper & Row, New York (1958)

Wittgenstein, L.: Philosophical Investigations (Trans: G.E.M. Anscombe, P.M.S. Hacker, & J. Schulte; revised fourth edition by P.M.S. Hacker & J. Schulte). Wiley-Blackwell, Chichester (2009)