The Philosophising Machine – a Specification of the Turing Test

Arthur C. Schwaninger¹

Received: 15 July 2020 / Revised: 3 August 2021 / Accepted: 27 January 2022 /
Published online: 4 March 2022
© The Author(s) 2022

Abstract

Block's (*The Philosophical Review*, 90(1), 5–43 1981) anti-behaviourist attack of the Turing Test not only illustrates that the test is a non-sufficient criterion for attributing thought; I suggest that it also exemplifies the limiting case of the more general concern that a machine which has access to enormous amounts of data can pass the Turing Test by simple symbol-manipulation techniques. If the answers to a human interrogator are entailed by the machines' data, the Turing Test offers no clear criterion to distinguish between a thinking machine and a machine that merely manipulates representations of words and sentences as it is found in contemporary Natural Language Processing models. This paper argues that properties about vagueness are accessible to any human-like thinker but do not normally display themselves in ordinary language use. Therefore, a machine that merely performs simple symbol manipulation from large amounts of previously acquired data – where this body of data does not contain facts about vagueness – will not be able to report on these properties. Conversely, a machine that has the capacity to *think* would be able to report on these properties. I argue that we can exploit this fact to establish a sufficient criterion of thought. The criterion is a specification of some of the questions that, as I explain, should be asked by the interrogator in a Turing Test situation.

Keywords Turing test · Natural Language Processing · Sorites paradox · Vagueness · Intelligence · GPT-3

1 Introduction

In order to make a well-founded judgement about the intelligence of a machine, Turing (1950) presents what is known today as the Turing Test (TT). The setup of the test can be roughly described as follows. A human interrogator is in a room

✉ Arthur C. Schwaninger
arthur.schwaninger@uzh.ch

¹ Department of Philosophy, Theoretical Philosophy, University of Zurich, Zürichbergstrasse 43, Zurich 8044, Switzerland

where she can send text messages over a digital device to two communicators, one of which is a machine. By asking questions and receiving answers, it is the interrogator's task to distinguish which of the two communicators is a human and which one is a machine. It is the objective of the machine to mislead the interrogator into making a wrong assessment by imitating human response behaviour. If multiple interrogators in repeated trial-experiments fail to make the right assessment, then the machine is said to have passed the TT which, according to Turing, means that the machine can be said to be intelligent. Turing does not, however, specify the precise success rate that is required to pass the test.

Turing's motivation for testing the behaviour rather than analysing the system's underlying thought-generating processes is that he believes we ought to avoid endless rehashing of intuitions about thought and replace questions about thinking with questions about the imitation game. As sometimes falsely suggested (e.g., French, 1990), Turing does not claim that passing the TT is necessary for a machine to be intelligent. Instead, passing the TT provides sufficiently good evidence that a machine has the ability of having thoughts. We can, for instance, imagine that intelligence manifests itself in forms that fail to imitate human life or language.

Today, Natural Language Processing (NLP) is a discipline dedicated to the development of computational language models that are trained to predict strings in order to generate words and sentences. NLP models are somewhat successful in their predictions of text strings as they learn to identify language's structure and usage from huge amounts of data. Following today's developments in NLP, one might be concerned about whether the TT really achieves what Turing argues. In his famous 'Blockhead' thought experiment, Block (1981) shows that the TT is not a *logically* sufficient criterion. Even though Block's thought experiment is physically infeasible, I argue below that similar concerns do in fact arise for practically relevant cases we might encounter in future NLP models. Even if a machine achieves back and forth communication with a human being, the original TT provides little evidence that can distinguish a thinking machine from one that is merely moving around words and using sentences it finds in an enormous database of texts or in a compressed form of representation. Even if the machine does not contain *all* possible conversational data between two communicating agents, it might be programmed to employ simple symbol manipulations and induction techniques on the text-based data that are capable to fool the interrogator even though it contains no thoughts at all.

With the rise of modern machine learning, this worry is not unjustified. Consider, for instance, the language model developed recently by the company OpenAI LP to generate coherent paragraphs of text from a given human input. An example of their results is the following (Radford et al., 2019a, b):

Human input:

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Short extract from the machine-written output:

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez. Pérez and his friends were astonished to see the unicorn herd

Even though this extract is merely the best result of numerous trials and even though this model was not specifically designed to interact in human conversations but rather to generate coherent, human-like texts, it does illustrate how powerful state-of-the-art NLP models can be and what we are to expect from the future. The power of OpenAI's algorithm does not come from a capacity of having thought but rather from the fact that the model was built on the basis of huge amounts of data, such that simple computational techniques on the compressed representation of this data allow the generation of human-like text outputs. As I explain in more detail below, modern NLP techniques might have a greater chance in succeeding in the original TT than we have imagined so far, even though they clearly have no capacity of generating thoughts. Henceforth, I will refer to the worry that the original TT might fail to distinguish between a machine that merely manipulates representations of words and sentences and one that is capable of having genuine thought as the *word manipulation worry* (WMW).

The TT is unspecific about how the interrogation process operates. Together with the above-mentioned worry, a strategy recommendation for the original TT is needed that specifies the test's procedure and clarifies what sort of questions the interrogation must ask to circumvent this worry. In this paper, I argue for a certain procedure of interrogation that enables the interrogator to differentiate between a machine that generates answers without thinking but merely by manipulating text data and a machine that cannot be faulted on this score. For the purpose of abbreviation, I will henceforth simply say that a machine that passes the proposed extended version of the TT can *think*, even though the ability to think might require further strategy recommendations to the original TT than those I present here. It must also be noted that the TT, including the proposed specification of the questions, merely captures an anthropocentric understanding of *thinking*. Other possible forms of thought are excluded.¹

The rest of the paper is structured as follows. In Section 2, I discuss the implications of Block's thought experiment and draw some conclusions relevant to how the

¹ As French (1990) correctly argues, the TT relies on an anthropomorphic notion of intelligence and it ultimately answers the question whether the machine is capable of having human intellect.

TT must be conducted. In Section 3, I explain why the test cannot be purely behaviouristic; rather, we must at least know the kind of training data used to develop the system being assessed. I continue to give more details on what it means to *think* in Section 4 and specify the conditions for an adequate question that should be asked by the interrogator. I then argue, in Section 5, that the concept of vagueness will always remain opaque to an NLP model and that a question testing the understanding of the sorites paradox can distinguish between a thinking and a non-thinking system. I conclude by presenting a slightly altered version of the TT that specifies the interrogator's questions in order to avoid the WMW and sets certain constraints on the data by which the system was developed.

2 NLP and its Reliance on Vast Amounts of Data

An example of a theoretical machine that passes the TT while arguably not possessing any thoughts by itself was proposed by Block (1981). The envisioned machine – called the *Blockhead* – is programmed using a huge look-up table that is organised in a tree structure whose first level contains every sentence (or group of sentences) that could possibly be the opening sentence in a conversation. The second level contains every possible reply to all these sentences at the first level. The third level contains any possible response to the second level and so on. Ultimately, the table contains all possible discussions that the machine could have with a human interrogator in a finite time span. As the Blockhead could be programmed using finite resources, it is logically (though not physically) possible for there to exist such a machine.

Some might argue that this type of machine does in fact think. However, it is more plausible that the capacity of having thoughts was *required to build* the machine, rather than the machine actually having this capacity itself. While Block does not explain how the vast dataset of the Blockhead could come into being, the sentences encoded in the tree structure of Block's machine depend on a thinking being in the same way as books contain sentences that require a thinking being. The machine is in no way causally involved in generating these thoughts. It merely transmits them from one or multiple thinking agents to another.

The Blockhead thought experiment is intended to illustrate that merely evaluating the behaviour of a machine does not suffice to judge its intelligence – at least from a logical perspective; thus, some knowledge about how the system was developed is required. However, just as importantly, the thought experiment also shows that the more data there is available, the greater the chance of passing the TT. If we were to remove a single leaf from the Blockhead's look-up table, the probability of failing the TT would still be virtually zero. Also, assuming that certain conversations are much more likely than others, the removal of all unlikely conversations would not necessarily lead to a low TT success rate. The (text-based) data, e.g. large collections of gathered texts such as the conversational texts within the Blockhead look-up table's tree structure, will henceforth be referred to as 'training data' (TD). TD is the part of the program that can be altered without changing the program's logic, i.e. the conditional clauses or control flow statements. The term is used here more broadly

than ordinarily used in machine learning, as it will also include any hard-coded sentences in a program as in the case of the Blockhead.

In some sense, the Blockhead's tree-model can be seen as a limiting case of contemporary big data models in NLP. The performance in modern machine learning algorithms employed in the development of text generating machines heavily depends on the amount of data available for developing the computational model. The use of conceptually simple machine learning architectures is compensated for by the employment of enormous amounts of data in order to train the model. Surely, in contrast to the Blockhead, this data is usually not explicitly stored within the computational model. Nonetheless, the computational linguist's hope is that by providing the model with enough data, it will eventually capture the necessary structures within language in order to build a model that is capable of performing linguistic tasks at a level that compares to humans. Consider the example already touched upon in the introduction of this paper: The cutting-edge text generation algorithm from OpenAI was trained on a dataset of 8 million web pages (40 GB of internet text), resulting in a model that 'feels close to human quality and shows coherence over a page or more of text' (Radford et al., 2019a, b). The developers claim that their model is so successful in capturing human parlance because their system pulls from a much greater amount of data, and from a larger variety of text-types, than what is commonly employed (Radford et al., 2019a, b).

As McDermott (2014) explains, it is true that the size of the data required to make Blockhead work is so preposterously large that Block had to violate the laws of physics to get it to fit in the known universe. However, with regards to data compression, the Blockhead is also one of the worst possible algorithms one can think of to succeed in the TT, as it has all its data stored *explicitly* in memory. It is far less clear whether the learning algorithms from NLP might not be able to compress and extract information from a *physically feasible* amount of data to such an extent that it might be equally successful. Nowadays, successful text generating algorithms are successful largely because they have been trained on so much data that simple techniques of inference and text manipulation will lead to good results. What I here describe as 'simple techniques' are the ordinary techniques already used in NLP today – or, at least, nothing far from it. (Section 4.1. explains in detail what these models *cannot* do.) The data does not have to be represented in explicit form, as in the case of the Blockhead where every sentence is stated explicitly in the tree-structure. Instead, data is usually compressed in terms of some representation that captures the important structures and features of the data by generalisation.² The sentences are not compressed in the case of the Blockhead, but realistic systems might nonetheless achieve similar levels of performance by learning from huge – but still

² Many modern systems encode text-based data as relatively low-dimensional representations, for instance, by the use of *word embeddings* and *document embeddings* (Bengio et al., 2003; Koehn, 2017, p. 37ff; Widdows, 2004). Word embeddings capture the semantics of a specific word in the training data by mapping it to a word vector, thereby capturing a low-dimensional representation of its features. Document embeddings capture the semantics of a whole sentence or document in the training data by mapping it to a similar vector space. Both allow for the execution of semantic calculations being performed over the corresponding vector space (Mikolov et al., 2013).

feasible – amounts of TD, compressing this data and decoding the compressed representation on demand. On my interpretation, the Blockhead implies that a machine which is successful in the TT by performing simple inference on enormous amounts of the represented TD cannot be trusted to contain thoughts. This is the WMW.

3 Controlling the Machine's Training Data

My analysis so far shows that the performance of any machine in a TT will depend on the TD with which it was developed. As a heuristic, the more TD is supplied, the higher the probability of succeeding in the TT. If a system was developed with barely any TD, it will obviously fail the TT, no matter how intelligent it might otherwise be. If a system possesses, explicitly or implicitly, a lot or even all possible finite conversational data, it can, even without sophisticated computations, pass the TT. WMW can then be restated in terms of principle P-1:

(P-1) During a TT, any answer A that the machine gives to a question Q cannot reliably be said to provide evidence about the machine's capacity to think, because the answer may have been delivered from manipulating and performing simple inference techniques on the representation of the TD and the question.

Thus, testing the machine on TD that were put into the system at its time of development or that were collected by the machine prior to the experiment does not lead to much valuable information about whether it can think. For my purposes, this is the relevant insight to draw from Block's thought experiment.

Conducting a behaviourist test for the purpose of finding out whether a machine can think is sensible in so far as there is little hope to draw similar conclusions merely by dissecting and analysing the machine's program or model. This is made apparent by the struggle researchers have today when having to explain why a machine learning model makes one decision rather than another. It is thus for good reason that almost any modern learning model (particularly one which relies on multi-layered neural network computations) is said to be a 'black box,' epistemically speaking. However, even if we cannot understand the machine's reasons for saying something or for drawing certain conclusions, we are always in a position to tell what TD was used in the development process. This holds no matter where technology will lead us in the future. The TD is the starting point from which the machine draws its conclusions and thoughts, if it has any. Even if a machine is trained using images, video, sound or any other form of data, the data can be available – if the machine's developers try to do so – simply because it is a machine – rather than a biological system – that by definition receives data in a digital format that could in principle be stored. Furthermore, as the TD, by definition, does not come from *within* the machine program's logical structure but from *outside*, it can always be modified. The next principle, P-2, follows:

(P-2) For any machine, full knowledge of the TD used for developing the machine is possible and the TD can be altered.

Even though we might have no idea how a machine operates, all that (P-2) is saying is that we are at least in full control over the data available to the machine; we always know *on what basis* it operates. In this sense, we do know at least something that goes beyond the mere behaviour of the machine. So even though I treat the machine as a ‘black box,’ I do assume that we know more about it than merely its behaviour.

4 Requirements on the Test

4.1 Having Intelligent Thoughts

Turing believed that any machine that could talk as if it was having thoughts (or sensations, or any other mental phenomena) would be just as difficult to build and interesting to study as one that really did have thoughts (or sensations, etc.). Although he motivates the TT as a replacement for any debates about what it means to ‘think,’ he himself does not stick to his replacement idea throughout his *Mind* paper as he discusses various different topics. The Blockhead thought experiment and the WTW also require us to engage in a discussion on what it means to have thoughts.

Bender and Koller (2020) argue that the structure or form of language does not reveal its meaning, because language patterns alone provide no information about the communicative intents of a text. Rather, the intents lie outside of language, within the ‘world.’ A system must not merely detect patterns in the data but must rather be able to connect the data to the world – or, as many others would say, judgements of the text must be based on how the text’s meaning is *grounded* to the dynamics of the environment (Harnad, 1990). Thus, an NLP language model would be incapable of having an *understanding* of what it is saying. It is only due to our *active participation* as listeners that we find meaning in the strings a machine returns. Accordingly, Bender and Koller correctly urge us to be ‘extra careful in devising evaluations for machine understanding’ (p. 5188). They present a thought experiment on the basis of which they argue that an NLP model will not pass a weak form of the TT if confronted with a new situation that requires an adequate understanding of how language relates to the world. The thought experiment goes roughly as follows: Two isolated human agents A and B communicate with each other via a cable at a distance. A system, O, that can detect statistical patterns (e.g. an NLP language model) follows the conversation unnoticed and eventually manages to predict B’s response to A’s messages. At some point, O cuts the wire and tries to imitate B without A noticing the difference. According to Bender and Koller, as long as the A and O are engaged in making utterances that have a social function, A will not notice any difference. However, at some point A requires O’s help that depends on having novel ideas about how one could align certain objects in the world in order to solve a certain problem. It is at this point where Bender and Koller believe that the machine can no longer fool A. According to them, ‘people constantly generate new communicative intents to talk about their constantly evolving inner and outer worlds’ (p. 5193) and an NLP model fails to deal with such situations.

Bender and Koller's argument won the 'Best Paper Award' at one of today's most prestigious NLP conferences. Nonetheless, one assumption in their argument does not seem to fit into actual NLP research. The person interacting with the NLP model is already familiar with all the data on which the model was trained; thus, she is no longer fooled by the model when she asks a question which she cannot answer herself and which demands a new way of thinking about some actuality in the world. The thought experiment must assume that A's question is disconnected from everything A and B have previously talked about (otherwise, A would easily come up with an answer himself). However, in NLP research, it is ordinarily much less clear whether the answer to a question required thought and understanding or whether it is subject to the WMW. We might know a bit about what kind of text jargon has been used as TD but not precisely what can and cannot follow from these texts by simple word manipulation. On the one hand, the machine must have *some* acquaintance with the terms involved or else it cannot even have a most basic conversation. On the other hand, to achieve any well performing system with roughly today's NLP techniques, the amount of text used as TD typically exceeds the amount of text any human has ever read in his or her lifetime. This makes it impossible to know whether the TD already contains so much information that the NLP model can come up with an adequate answer by simple word manipulation or whether the model requires the capacity of having thoughts in order to come up with an answer.

As Smith (2019, ch. 6) discusses in his recent book, we see that modern systems fail to ground meaning in the world by the fact that they fail to distinguish between correlating and causal factors and are thus famously prone to making racist judgements and typically fail to distinguish false from accurate news reports. For him, a crucial point about genuine intelligence is the ability to know what one is talking *about*. So, an intelligent system 'has to be able to *distinguish appearance from reality*. ... The system must recognise ... that the object [of the world] is *different* from its representation of it' (Smith, 2019, pp. 84–85). The system must also be committed to the identified objects and their categories and must *care* about the difference these categories make. This allows the system to distinguish right and wrong suppositions about the world. We roughly summarise this aspect of Smith's proposal in terms of a single principle:

(P-3) To have genuine intelligence, a system must understand what object it is talking *about* and must distinguish its representations from the object itself.

To be sure, the concerns raised against standard approaches in NLP are not concerns that should make us doubt that human brains are operating as symbol-manipulation machines. Although humans display intelligent behaviour by relying on a brain that performs some form of computation, the way NLP models learn and operate seems to be fundamentally different. Most importantly, the data by which these models are trained are purely text-based. They do not contain any visual data nor data that connects actions with sensory input. While the human brain might truly be only manipulating symbols, these computations are in some relationship to how the agent acts and stands in the world. One of the essential insights of *embodied cognition* is that cognitive processes are not exclusively located in the nervous system but depend

on some relationship between the agent and the world.³ Likewise, the suspicion here is that having genuine thoughts about the world might require computations that are related to the world and to actions in the world.

While Smith would probably deny that contemporary NLP techniques or anything like them in the future could ever pass the TT, I think differently. We should not underestimate the seemingly emerging abilities that come about once a lot of data is used for the development of an NLP model. Consider, for instance, OpenAI's most recent language model called GPT-3 (Brown et al., 2020), which has been said to show 'strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic' (p. 1). In particular, the developers claim that their model has the ability of mathematical reasoning without task-specific training (pp. 21–23). The model was asked 'What is 48 plus 76?' which the model managed to answer correctly by outputting the string-token '124.' The developers do not know exactly how the model does this, nor could we have expected these results based on pure reasoning. Instead, GPT-3 possesses seemingly human-like reasoning techniques as a result of employing billions of parameters to represent the structure of the roughly 45 TB of text data.

Identifying the limits of NLP is extremely hard and philosophers can become too self-assured in their assessment. For instance, take Dennett (1998), who draws attention to the Winograd-Task⁴, where a computing system must correctly identify the reference of a pronoun in each of two sentences that differ by only one word. For example:

'The committee denied the group a parade permit because they advocated violence.'

and

'The committee denied the group a parade permit because they feared violence.'

Here the pronoun 'they' is grammatically ambiguous but semantically unambiguous to humans. Dennett's assessment of this task is as follows:

[I]f sentences like this are embedded in a conversation, the computer must figure out which reading of the pronoun is meant, if it is to respond intelligently. But mere rules of grammar or vocabulary will not fix the right reading. What fixes the right reading for us is knowledge about the world, about politics, social circumstances, committees and their attitudes, groups that want to parade, how they tend to behave and the like. One must know about the world, in short, to make sense of such a sentence. (p. 38)

³ Central ideas of embodied cognition are already discussed by Gibson (1966).

⁴ The challenge was first formulated by Winograd (1972).

However, the developments in NLP seem to indicate that Dennett is wrong. While state-of-the-art NLP models sometimes achieve a success rate of around 80%⁵ in such tasks, GPT-3 also achieves rates above 70% and I imagine that future models may perform even better without having to revolutionise contemporary NLP techniques.

All in all, I do believe that contemporary modelling techniques used in NLP do not provide models that are capable of having thought. However, I can imagine that a physically feasible, non-thinking system may someday perform very well in the TT if we do not have *very* well worked out questions for the machine to answer. I therefore want to make a suggestion as to what the interrogator should ask during the TT in order to make a competent judgement.

4.2 Formulating a Question

Assuming there is a machine that performs well in the TT, then, in light of the WMW, the TT could be extended by a recommendation to the interrogator to ask a precise question which demands an answer that cannot be directly inferred from the represented data but is, in this sense, *novel*. Additionally, (P-3) demands that the question requires an understanding that distinguishes world from representation.

The challenge, therefore, is to come up with a question (possibly including follow-up questions) that satisfies the following conditions:

- C1. The question has an answer which could only by chance be generated by simple text manipulation and inference techniques on the data of the question and the TD taken together.
- C2. The question is not too hard to answer, as harder questions increase the number of false-positive test cases. (The harder the question becomes, the harder it becomes for any thinking system to answer it correctly.)
- C3. There is a fairly precise criterion for verifying the success of the machine.
- C4. The answer does not depend on any facts about the world outside of the TT environment, aside from the facts already logically entailed in the TD and the question itself. Otherwise, the question would be impossible to answer correctly without merely guessing.
- C5. Nonetheless, the question should require the system to distinguish between representation and reality and to prove that it can tell that the distinction is important.

According to principle (P-2), one has full control over the data in the TD. Thus, in order to satisfy C1, the question we are looking for determines how we alter the TD such that the answer to the question is not merely entailed. However, the modification of the TD should not have an effect on the machine's performance during an ordinary conversation in the TT. Of course, it should not be possible to answer

⁵ According to Sakaguchi et al. (2019) human performance on the task is 94.0%.

the question by drawing inferences from the question together with the TD. For the purpose of brevity, I will henceforth assume the interrogator's question to be part of the TD.

One might be inclined to formulate two questions, where in one case all conditions except C5 are satisfied and in the other case all conditions except C4 are satisfied. This would, however, give rise to the worry that the different aspects of thought are possibly not connected to one another. Having a single question that requires self-aware, world-oriented thinking (C5) that is not fact-based (C4) is therefore considered the better strategy.

4.3 The Philosophising Machine

While some of the sciences, for instance cognitive psychology, have an interest in understanding the relationship between the world and human thought, it is philosophy that is most strongly engaged in tying together subjective thought and its relation to the world, sometimes without relying on any evidence other than subjective mental thoughts themselves. The armchair philosopher frequently draws her intuitions from contemplating thought and its relationship to the world.⁶ Even though NLP algorithms might someday be able to imitate human speech under almost all ordinary circumstances, the capacity for much of what counts as philosophical talk is distinctive to thinking beings, because philosophising of this sort depends on the capacities discussed in Section 4.1 that, according to Smith, sufficiently determine intelligent thought.

All that an NLP model is capable of doing is to *use* language consistently and informatively. But it will never, in NLP's current form or anything similar to its current form, be able to identify and contemplate what it is subjectively doing when using language. These deficiencies can, or so I argue, be exposed in philosophical discussions where no prior knowledge or linguistic competence is required. The question now is how, in a TT situation, we can force a machine to make philosophical reflections and notice if it has actually done so without setting a challenge that is too difficult even for ordinary humans.

In order to find a good question which satisfies C1–C5, one could, for instance, suggest removing all philosophical texts from the TD and see whether a machine could come up with such thoughts by itself. One could think of testing the machine on some synthetic, *a priori* truth (if one is not opposed to these Kantian notions) that cannot be warranted by inference from some preselected TD. For Kant, these truths include mathematical facts and the truths of metaphysics. However, neither of these two domains fulfils all of conditions C1–C5. Metaphysical truths depend largely on what one's metametaphysical conceptions are, which again depend on having a rich understanding of the subject matter. Answers to metaphysical questions are also hard to evaluate as there is seemingly little consensus on any question. Furthermore, the problem might be too hard such that the machine might, for

⁶ This is not to say that philosophy works *exclusively* on intuitions but it traditionally takes manifest experiences as the starting point of its reasoning.

understandable reasons, answer ‘I don’t know’ and could not be judged on these grounds. Mathematical truths suffer from similar problems and in addition are not clearly coupled to the world; so, many questions about mathematics will violate C5. This is one reason why in fact computers are astonishingly helpful to humans in finding new proofs.⁷

Expecting a philosophical theory from a machine or a human will not bring us very far. However, this is also not what we really need to make our assessment. Rather, we must merely find evidence that the machine *understands* the philosophical problem. It does not need to come up with a solution. I can imagine that there are numerous philosophical problems of this sort and I leave it to the reader to come up with examples other than the one I am presenting now.

5 Vagueness

There are certain logical or semantic paradoxes that are perplexing in terms of reference or denotation, the relation between a referring expression and its referent. These include, for instance, the *Liar* as well as *Berry’s paradox*. Formulating a question on their behave might satisfy conditions C1 to C4, but not C5. There is, however, one paradox that is partially semantic and partially metaphysical, i.e. directed to reality. I am thinking of the *sorites paradox* that runs as follows. One grain of sand is not a heap. Adding another grain of sand to what is not a heap does not make it a heap (induction step). From this it can be logically inferred that a million grains of sand are not a heap. Despite the conclusion being obviously false, both premises appear to be true, giving rise to a paradox. The induction step seems intuitively true because a tiny grain of sand cannot possibly make the difference between being a heap and not being a heap. In logical terms, the paradox can be formulated in the following way:

$$\begin{aligned} \text{P1: } & Fx_1 \\ \text{P2: } & \forall n (Fx_n \rightarrow Fx_{n+1}) \\ \text{C: } & \forall n Fx_n \end{aligned}$$

Here F represents a predicate such as “not a heap” and x_i refers to the i -th object in a sorites sequence – in our example, a pile of i grains of sand.

The sorites is intimately related to the phenomenon of *vagueness*. Arguably, vagueness can be reduced to two characteristic symptoms (Greenough, 2017). First,

⁷ Automated theorem proving was already possible in the 1950s. Back then, Newell and Simon (1956) developed a program that relied on symbol-manipulation techniques to construct proofs of some of the theorems in Russell’s and Whitehead’s *Principia Mathematica*. In a letter to Russell, Simon summarises their findings:

We denied ourselves devices like the deduction theorem and systematic decision procedures of an algorithmic sort; for our aim was to simulate as closely as possible the processes employed by humans when systematic procedures are unavailable and the solution of the problem involves genuine “discovery.” ... [We] find that the proofs it produces resemble closely those in *Principia*. (Simon, 1996, p. 208–209)

there appear to be cases x_i for which we do not know whether Fx_i or $\neg Fx_i$. These are called *borderline* cases. Second, the predicate's extension appears to lack sharp boundaries such that a small change of x_i , like the addition of one grain of sand, does not appear to affect the truth value of the statement Fx_i .

As Shapiro (2010) explains, 'the linguistic and worldly components to vagueness are thoroughly intertwined and cannot be disentangled' (p. 161). Human thought arguably relies on the conceptualisation of the world by the means of vague terms.⁸ It is to a large extent not up to us whether we want to conceptualise the world in vague terms. Rather it is our unconscious thought process that gives us no other choice than to do so. As Wright (1987) explains, to communicate effectively in social groups our minds have evolved in a way that a certain *tolerance* of certain predicates is ensured, such that small differences about things in the world do not matter for predication.

Philosophical views might not have fully converged on this subject matter; however, it is fair to say that part of *understanding* the sorites lies in understanding that the representation of an object, expressed by the vague predicate F , is distinct from the actual object. If this distinction were not to be understood by a system, then the system would also not understand the force of the paradox. For instance, an NLP model might well be able to correctly manipulate vague terms such as 'heap' and output sentences that correctly relate 'heap' to 'a grain of sand' because it has learned to *use* these terms effectively. The model has learned from the TD how to *use* language in a way that they might potentially fool an interrogator during a TT. Yet, the mere *use* of vague terms does not ordinarily reveal the properties of vagueness – assuming all philosophical and logical texts on vagueness were removed from the TD. In contrast to any such NLP model, humans *do* understand the paradox rather quickly, not because of their capacity to *use* language but because of their capacity of reflection and thought. Even though most human non-philosophers have never heard about vagueness they can fairly quickly *understand* the paradox in the sense that they feel the seductiveness of the induction step and understand that vague terms and the objects they refer to have a somewhat strained relationship.

If we accept that an understanding of vagueness cannot be imitated by an NLP model, we should advise the interrogator during the TT to test the two communicators in such a way that he can tell when one of them does not seem to grasp the concept of vagueness. Of course, regarding the test's setup, we must at least assume, per (P-2), that we know the machine has not been trained on texts about vagueness. In the following, a machine that was trained without any data containing information about logic, vagueness or philosophy is indicated with the asterisk symbol. Aside from matters of convenience, this emphasises that the test is not purely behaviouristic. While the scientists conducting the test do not know what their created 'black box' is actually doing, they do know the TD on which its behaviour is based, which is an acceptable assumption, according to (P-2).

⁸ Philosophers have tried to pinpoint the precise psychological reasons behind the phenomenological sorites paradoxes (Raffman, 1994).

6 Specifying the Original Turing Test

Given my reasoning above, what question could be asked to a machine* in order to find out if it has human-like thoughts? I believe, one would simply have to present it with the sorites argument and ask, ‘Is there something wrong with this argument?’ Different options of responses exist here. First, if it says, ‘No,’ then the machine* has obviously failed the test. Second, if it says, ‘I don’t know,’ then the interrogator must push further to get an answer, by asking why it does not know or even by saying, ‘I’m astonished that you do not know, normally people do know,’ to put further pressure onto the machine. Third, if the machine* says, ‘Yes,’ or, continuing the second case above, changes its opinion to, ‘Yes,’ then one continues by asking, ‘Why?’ Now, we might be inclined to think that the machine* will at this point try to talk itself out of the problem by saying odd things like, ‘No one ever adds one grain to a bunch of sand,’ or something like, ‘I hate it when philosophers try to trip me up this way. I’ll take a zero on this question and we can go on to something else.’ I would find this behaviour to be already slightly suspicious, because I would expect most people to at least acknowledge that something dubious is going on in the formulation of the induction step. But in any case, I argue that a machine* would never say anything of this sort. Assuming that it would try to talk itself out of the situation already assumes that it is *aware* that this is a somewhat exceptional question that should be avoided. But for the machine* to have this awareness, the machine* would already have to have a true understanding of the problem, which is what we are testing for.

Consequently, it is the lack of hesitation which should make us suspicious about the response of a non-thinking machine* because it would show that the machine* finds the problem trivial. Finding the problem trivial means that the machine* does not find the induction step seductive enough, which, I contend, shows the lack of having thought. Alternatively, it could be that the machine* simply finds the problem trivial due to the fact that its intelligence surpasses the thought capacities of any human being by many orders of magnitude, and that it failed to adapt itself accordingly to the situation because it did not have any information about how humans respond to such problems. But in this case, the machine can also provide us with good reasons as to why it finds the induction-step to be false, and having these reasons requires having thought.

I believe that (the lack of) hesitation will already be enough to distinguish a machine that is merely manipulating word representations and one that is capable of having thought. However, to be certain in our assessment, we can ask the machine* to explain its *reasons* for believing that the second premise is false.

At this point, there seems to be an additional problem in that the solutions to the paradox are so vast that the machine could say almost anything, and it would be in line with some solution. For instance, the machine* might express the view that there really is a sharp boundary between heaps and non-heaps along a sorites sequence, as supported by Williamson’s (1994) epistemic theory of vagueness. If one wants to resist the temptation of casting doubt upon Williamson’s capacity of thought, one must specify a criterion that determines whether or not the machine*

has managed to think about the properties of vagueness, *independent* of its opinion. Luckily, there is one thing that all philosophers can agree on, and that is that a successful theory ought to give an answer to what Fara calls the *psychological question*:

If the [sorites' induction step] is not true, why were we so inclined to accept it in the first place? In other words, what is it about vague predicates that make them seem tolerant and hence boundaryless to us? (Fara, 2000, p. 50)

The psychological question tries to identify the *reasons* why we are inclined to accept the induction step even if it is false. For instance, Williamson would argue that our intuitions are formed by the fact that we do not know where the cut-off point is. If there were a cut-off point somewhere along the sorites sequence running from F and $\neg F$, then we as competent speakers would expect that we would also know where it is. Hence, as we do not know where the cut-off point is, we are falsely inclined to accept that there is no such point. This makes us more likely to accept the sorites' second premise.

So, even if the machine* is only somewhat hesitant about its judgement regarding the second premise, or even if it was not hesitant at all, if the machine* feels the need to answer the psychological question, we know that it has understood the concept of vagueness, because it has exhibited the ability to contemplate its own thoughts and their relation to the world without making inferences from the TD. We must not expect the machine to give any right or truly consistent answer, but we can expect it to show awareness of the problem if we give it enough time to think about it. It might be that the machine* fails to give a convincing answer to the problem of the sorites. However, if a machine* does not even try to answer the psychological question, one should be in doubt about whether it has actually understood the problem at hand because it might be merely manipulating words and sentences. If, however, the machine* does make attempts to answer the question, or proves to be aware of the problem, one has sufficient reason to believe that the machine* can think – at least in the sense and to the extent specified earlier.

As the psychological question is addressed by any somewhat reasonable theory of vagueness, addressing or pointing out the problem serves as a criterion that seems easy enough to be fulfilled by any thinking machine. I have therefore developed a sufficient criterion to distinguish between thinking machines and machines that merely manipulate representations of words and sentences, thereby avoiding the WMW. I believe this criterion would be easily fulfilled by a human-like thinking machine and does not require the possession of any specific knowledge. The proposed criterion therefore fulfils the above conditions C1–C5.

7 Conclusions

The Blockhead thought experiment reveals that the amount of TD used to develop a machine impacts the success at passing the TT. This is exemplified by modern NLP systems that demonstrate an astonishing ability to imitate humans merely by performing simple inference and text manipulation techniques on their representation of huge amounts of TD. Such systems are argued to not possess any thoughts by

an anthropocentric notion of the term. Instead, I agree with others that a necessary condition for systems to have thoughts is that they rely on an information-processing machinery which involves computations that connect symbols and computational states with the world.

To test whether a system might potentially have thoughts, I have argued that one performs a TT which involves questions about vagueness. Since the TD merely contains information about the *use* of words, the answer to a question about vagueness cannot be derivable from the TD if that data had been filtered accordingly. I have argued that my criterion avoids a realistic worry we might have about a machine that proves to be successful in an ordinary TT situation.⁹

Funding Open access funding provided by University of Zurich. This research was possible thanks to the Forschungskredit of the University of Zurich, grant no. FK-19-065.

Data Availability None.

Code Availability None.

Declarations

Conflicts of Interest None.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198
- Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155
- Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, 90(1), 5–43
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al. (2020). Language models are few-shot learners. *arXiv:2005.14165 [cs]*. <http://arxiv.org/abs/2005.14165>. Accessed 6 Jul 2020
- Dennett, D. (1998). Can machines think? Reprinted 2002. In D. J. Levitin (Ed.), *Foundations of Cognitive Psychology: Core readings*. MIT Press
- French, R. M. (1990). Subcognition and the limits of the Turing test. *Mind*, 99(393), 53–65
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Waveland Press, Inc

⁹ This research was possible thanks to the Forschungskredit of the University of Zurich, grant no. FK-19-065.

- Graff Fara, D. (2000). Shifting sands: an interest-relative theory of vagueness. *Philosophical Topics*, 28(1), 45–81
- Greenough, P. (2017). Neutralism and the Observational Sorites Paradox. In A. Abasenezhad & O. Beuno (Eds.), *On the Sorites Paradox* (p. 42). Springer (forthcoming)
- Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42, 335–346
- Koehn, P. (2017). Neural machine translation. *arXiv preprint arXiv:1709.07809*
- McDermott, D. (2014). On the claim that a table-lookup program could pass the Turing test. *Minds and Machines*, 24(2), 143–188
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*
- Newell, A., & Simon, H. A. (1956). The logic theory machine - A complex information processing system. *IRE Transactions on Information Theory*, 2(3), 61–79.
- Radford, A., Wu, J., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019a). Better language models and their implications. OpenAI. <https://openai.com/blog/better-language-models/>. Accessed 10 Jul 2020
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019b). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 24
- Raffman, D. (1994). Vagueness without paradox. *The Philosophical Review*, 103(1), 35
- Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2019). Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*
- Shapiro, S. (2010). Vagueness, metaphysics, and objectivity. In R. Dietz, & S. Moruzzi (Eds.), *Cuts and Clouds* (pp. 149–162). Oxford University Press
- Simon, H. A. (1996). *Models of My Life*. MIT Press
- Smith, B. C. (2019). *The promise of artificial intelligence: reckoning and judgment*. The MIT Press
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460
- Widdows, D. (2004). *Geometry and meaning* (773 vol.). CSLI Publications Stanford
- Williamson, T. (1994). *Vagueness*. Routledge
- Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, 3(1), 1–191
- Wright, C. (1987). Further reflections on the Sorites paradox. *Philosophical Topics*, 15(1), 227–290

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.