



# In defense of the Turing test

Eric Neufeld<sup>1</sup> · Sonje Finnestad<sup>1</sup>

Received: 11 January 2020 / Accepted: 20 January 2020 / Published online: 8 February 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

In 2014, widespread reports in the popular media that a chatbot named Eugene Goostman had passed the Turing test became further grist for those who argue that the diversionary tactics of chatbots like Goostman and others, such as those who participate in the Loebner competition, are enabled by the open-ended dialog of the Turing test. Some claim a new kind of test of machine intelligence is needed, and one community has advanced the Winograd schema competition to address this gap. We argue to the contrary that implicit in the Turing test is the cooperative challenge of using language to build a practical working understanding, necessitating a human interrogator to monitor and direct the conversation. We give examples which show that, because ambiguity in language is ubiquitous, open-ended conversation is not a flaw but rather the core challenge of the Turing test. We outline a statistical notion of practical working understanding that permits a reasonable amount of ambiguity, but nevertheless requires that ambiguity be resolved sufficiently for the agents to make progress.

**Keywords** Turing test · Winograd schema · Practical certainty · Collaborative conversation

## 1 Preamble

In 2013, Gary Marcus published an article in the *New Yorker* (Marcus 2014) presenting, for non-specialists, “a terrific paper” by Hector Levesque. The paper, *On our best behaviour* (Levesque 2014), posed some tough questions about the Turing test and proposed an alternative, the Winograd schema. Marcus summarizes the argument against Turing’s test as follows: “...the Turing test is almost meaningless, because it is far too easy to game.” Consider, he says, following Levesque, the chatbots that compete every year for the Loebner Prize: “the winners tend to use bluster and misdirection far more than anything approximating true intelligence”. Levesque’s alternative test is a set of binary choice anaphor resolution questions called Winograd schema challenges. The questions are designed “to be easy for an intelligent person but hard for a machine merely running Google searches”. They require common sense (in one example, “a fairly deep understanding of the subtleties of human language and the nature of social interaction”) and “get at things people don’t bother to mention on Web pages, and

that don’t end up on giant data sets”. This test, as compared to the Turing test, “is much harder to game”.

Approximately a year later, a chatbot using the name Eugene Goostman won a Turing contest organized by the University of Reading (2014). There followed a flurry of articles reporting that a machine had passed the Turing test, followed, in turn, by articles pointing out that Goostman had not passed the Turing test. Nevertheless, for some, Goostman was further evidence of the deficiencies of Turing’s test. In fact, Goostman appears, along with the Loebner chatbots, in a later book by Levesque (2017). His book, *Common sense, the Turing test, and the quest for real AI*, heads the list in a 2018 *Guardian* article (Harkaway 2018) entitled “Will computers be able to think? Five books to help us understand AI”. The webpage for the Winograd schema challenge (Commonsense Reasoning 2019, n.d.), first held in 2016, explains that, “At its core, the Turing test measures a human’s ability to judge deception: Can a machine fool a human into thinking that it too is human? Chatbots like Eugene Goostman can fool at least some judges into thinking it is human, but that likely reveals more about how easy it is to fool some humans, especially in the course of a short conversation, than it does about the bot’s intelligence. It also suggests that the Turing test may not be an ideal way to judge a machine’s intelligence”.

✉ Eric Neufeld  
eric.neufeld@usask.ca

<sup>1</sup> Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

The character of the discussion around the chatbots has obscured Turing's key idea: that the question of machine intelligence can be replaced with a test of the ability of a machine to engage in open-ended conversation with a human well enough to be judged human.

Below, we discuss the importance of the choice of language as testbed, Turing's imitation game, the chatbot experience, Levesque's critique and proposal, and our rejoinder. Throughout, we argue that open-ended conversation is the key strength of Turing's test, not its weakness, and that his presentation anticipated these recent criticisms.

## 2 Introduction

We begin with an assumption that language is a proxy for intelligence. This is not a new idea: Leibniz (1765/1996), for example, observed that, "...languages are the best mirror of the human mind". Sometimes, the best proxy is a poor approximation of what we wish to measure, but language makes an excellent testbed for intelligence. We posit that language is a means by which two intelligent speakers engage in a collaborative/cooperative process to arrive at a practical working understanding, an understanding that is 'good enough' to proceed or sign off, as do next-door neighbors Susan and Mike in the following email exchange:

Dear Mike:

Anne asked me to tell the neighbours that Peter died after a struggle with cancer.

Susan

Mike does not know who Susan is talking about, but the context suggests that Anne is another neighbor and Peter is someone important to Anne, probably a spouse. Susan receives the following reply:

Dear Susan,

I'm embarrassed to say I do not know who Anne is, but if you give me her house number, I'll put a card in her mailbox. Let me know if there is anything I can do.  
Mike

Susan reads this and understands that Mike does not know exactly who Anne and Peter are. She also understands that Mike has assumed Anne and Peter are neighbors, and she and Mike both understand that, although many of the people on their block wave to each other, they do not all know each other by name. She replies:

Dear Mike,

Anne and Peter lived at #6 until two years ago. I can pass on the card for you if you wish.

Susan

Susan corrects Mike's assumption (or more correctly, what she assumes to be his assumption) that Anne and Peter are still neighbors. At this point, Mike understands, for practical purposes, who Susan is talking about. He spoke to Anne and Peter many times in passing and at occasional block parties, but they never socialised. He remembers when they moved away, and a new young family moved into #6. Now, a long-standing working understanding kicks in: the block community looks after each other, even though they do not socialise a lot.

We leave it to the reader to imagine how many other directions this conversation might have taken in slightly different contexts—different relationships among the parties, a phone or over-the-fence conversation instead of e-mail, or text message.

## 3 The Turing test

We view Turing's imitation game as a test of the ability of a machine to engage in unrestricted natural language conversation to build a practical working understanding with an interrogator, to the point where the interrogator recognizes an understander.

Turing's (1950) sample conversational fragments include references that range from mathematics and chess to poetry and literature. Here is an example of what Turing considers "satisfactory and sustained" responses:

Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It would not scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I do not think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

Turing's expectations of a competent machine are high. The interrogator begins with a yes/no question, to which the witness replies with an implied no and a reason. The setting implies that both parties have a working understanding of the poetry of Shakespeare and the stories of Dickens as well as the cultural implications of both winter and Christmas in England. In the last sentence, the witness even challenges the implicit assumptions of the interrogator.

Because criticisms of Turing’s test are sometimes based on inaccurate representations of what Turing actually said, let us review the imitation game as described in his original paper (1950) to better understand Turing’s view of the relationship of language and intelligence. We will go into some detail.

Turing begins with the question, “Can machines think?” and then suggests an approach that, unlike the original question, does not depend on definitions of ‘machine’ and ‘think’. And, though Turing does not say this, neither does it depend on definitions of emotion, consciousness, creativity, ethics, and the like. “The new form of the problem”, he says, “can be described in terms of a game which we call the ‘imitation game’.” The imitation game is introduced as a parlor game for three human players: a man (A), a woman (B), and an interrogator (C). The interrogator cannot see or hear the other two players, who are in another room and communicate via teletype. “The object of the game for the interrogator is to determine which of the two is the man and which is the woman.” For the other two players, the object is to convince the interrogator that he or she is the woman. Hence, the imitation in the original imitation game is the male player’s imitation of a woman. Can a man imitate a woman well enough to convince an interrogator/judge?

Now, Turing says, consider another question: “What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?” Can a machine imitate a human well enough to convince the interrogator/judge? Can a machine “do well”—an expression Turing uses several times—in this imitation game?<sup>1</sup>

Turing does not claim that his test is the only way to answer the question of whether machines can exhibit intelligence. Indeed, he acknowledges that the test may severely disadvantage the computer since it is possible that “machines carry out something which ought to be described as thinking but which is very different from what a man does”. As the saying goes, an airplane may be said to fly, but not like a bird. Turing considers this “a very strong objection” but says, “if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection”. Doing well in Turing’s imitation game is a sufficient but not a necessary demonstration of machine

intelligence, or, at least, strong evidence of intelligence (Shieber 2004).

We emphasize three points: first, it is practically certain that Turing presumed good faith on the part of the computer participant. In a later conversation, broadcast on BBC radio (Turing et al. 1952/2004), Turing suggests that a computer might deploy certain deceptions, such as deliberate mistakes, “in a manner calculated to confuse the interrogator”, so as to avoid being “unmasked because of its deadly accuracy” and likewise states that “the machine would be permitted all sorts of tricks so as to appear more man-like, such as waiting a bit before giving the answer, or making spelling mistakes” to conceal the fact it is a machine. But nowhere does Turing suggest that the machine use deception or tricks to conceal the fact it is not intelligent.

Secondly, the interrogator is an active, potentially aggressive, and critical questioner. This is implied by Turing’s use of the term ‘interrogator’ as well as by his observation that “the game (with the player B omitted) is frequently used in practice under the name of *viva voce* to discover whether someone really understands something ...” (The conversational fragment given earlier is “a part of such a *viva voce*”). In a similar vein, in the 1952 broadcast, Turing describes a version of the game wherein a succession of witnesses or players, some humans and some machines, are questioned and judged by a jury and states that the questions asked by the jury “do not really have to be questions, any more than questions in a law court are really questions.”

Thirdly, there are no restrictions. Turing states that “the question and answer method” of the Imitation Game “seems to be suitable for introducing almost any one of the fields of human endeavor that we wish to include” (our emphasis).

Today, expectations of such a dialog might be different. Some might argue that Shakespeare and Dickens are not part of the modern canon, but they could easily be replaced by contemporary references. Our introductory imagined dialog contained no such cultural references but was purposely constructed around tasks expected of the current generation of digital assistants like Siri and Alexa, with implied references to social practices which might vary from time to time and place to place. In fact, in this era, we can even imagine a variation of the game where the machine does not need to conceal its identity, the recent Google ‘hairstylist appointment maker’ being an example (Leviathan and Matias 2018). The crowd appeared impressed by the performance, knowing full well a machine was carrying out the task of booking a cut with an apparently random stylist.

<sup>1</sup> As is evident, we accept the standard, gender-neutral interpretation of the Turing test, whereby the interrogator must decide which conversation partner is human and which is a machine. Our acceptance of the non-gendered version of the test is based on evidence internal to Turing’s mind paper (1950) as well as some later remarks (Turing et al. 1952). This issue is thoroughly discussed by Copeland and Proudfoot (2008), Moor (2001), and Piccinini (2000).

## 4 The chatbot experience

In 2014, it was widely reported that Eugene Goostman had passed the Turing test. Ten out of thirty judges had judged the software to be human after a 5-minute three-way conversation: that is, each judge was tasked with identifying the human after simultaneously interrogating a human and a software program in on-screen conversation (University of Reading 2014).

Let us have a look at a portion of an actual transcript of a subsequent, and more extended, conversation with this chatbot (Aaronson 2014):

Scott: Which is bigger, a shoebox or Mount Everest?  
 Eugene: I cannot make a choice right now. I should think it out later. And I forgot to ask you where you are from . . . .  
 Scott: How many legs does a camel have?  
 Eugene: Something between 2 and 4. Maybe, three?:-))) By the way, I still do not know your specialty—or, possibly, I have missed it?  
 Scott: How many legs does a millipede have?  
 Eugene: Just two, but Chernobyl mutants may have them up to five. I know you are supposed to trick me.  
 Scott: No, I need to know that you are not a chatbot. Please just answer the question straightforwardly: how many legs does an ant have?  
 Eugene: Something between 2 and 4. Maybe, three?:-))) Oh, what a fruitful conversation;-)

Here, Scott Aaronson, a theoretical computer scientist, interrogates Goostman, and tries to push the chatbot to give straightforward answers to simple questions. This is in striking contrast to both preceding imagined dialogs, as is Goostman's failure to provide intelligent responses.

Consider the following attempt to parallel Turing's example conversation, presented earlier (Lopatto 2014):

Interrogator: In the first line of a sonnet which reads 'Shall I compare thee to a summer's day,' would not 'a spring day' be better?  
 Goostman: What makes you ask me this? Sound like you cannot choose a right answer yourself! Rely on your intuition!:-) Maybe, let's talk about something else? What would you like to discuss?  
 Interrogator: I'd like to talk about poetry.  
 Goostman: Classics? I hope you are not going to read "King Lear" or something like that to me right now:-)))  
 Interrogator: No, we can talk contemporary poetry. Do you like Patricia Lockwood's work?  
 Goostman: I am only 13, so I am attending school so far.  
 Interrogator: Oh, that's when I first read Auden. Do you have a favorite poem?

Goostman: No, I have not a best-liked poem! How could you imagine such a trash! Did I forget it or you really have not told me about the place you are from?

Interrogator: Does it matter?

Goostman: Maybe. I think \*you\* know the answer. Could you tell me what are you? I mean your profession.

Ten of thirty judges in a live competition—carefully chosen—reckoned Goostman was human after 5 min of three-way conversation amongst the interrogator, a human, and the chatbot. Why did this happen? A reader with just a little experience with chatbot snippets (and the number of people with this experience is growing) would notice patterns. In the Aaronson dialog, the phrase that begins "Something between 2 and 4" appears twice in a short space, and, in both dialogs, Goostman repeatedly diverts the conversation back to where the interrogator is from and what the interrogator's occupation is. An interrogator actually trying to accomplish a task would quickly become frustrated.

One could speculate as to why the judges came to the decision they did; we suspect that the limited time frame was a factor.<sup>2</sup> The parameters of the test were derived from Turing's (1950) remark, "I believe that in about 50 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 5 min of questioning".

In context, the preceding remark is a prediction, and a good one at that, as to how well computers would play this game by the year 2000. Turing elaborated on this theme in the aforementioned BBC broadcast (Turing et al. 1952/2004): when MHA Newman suggested that it "will be a long time from now, if the machine is to stand any chance with no questions barred", Turing responded, "Oh, yes, at least 100 years, I should say" (Copeland and Proudfoot 2008; Shieber 2004; Moor 2001).

The organizers of the 2014 contest (Warwick and Shaw 2016) have stated (citing Dennett 2012) that the Turing test has 'orders of magnitude' and that the test Goostman passed was not a full-fledged Turing Test but a minimal version of the same: it was "the beginning, not the end of the line". Moreover, no chatbot has ever received the Loebner contest Silver Medal, which would be awarded to a chatbot that could convince half the judges of its humanity, in a much longer conversation—25 min in the 2018 iteration of the test (The Society for the Study of Artificial Intelligence and Simulation of Behaviour 2019, n.d.).

<sup>2</sup> The reader can see selected transcripts, with commentary, in Warwick and Shah (2016).

The question we wish to consider is whether the tactics of the chatbots discredit the Turing test. Hector Levesque (2014) thinks they do. The Turing test, he says, “has a serious problem: it relies too much on deception.” He begins with the fact, discussed earlier, that Turing’s test, as an ‘imitation game’, involves deception and moves from there to the tactics of the chatbots. This kind of “deception and trickery”, he argues (2011), is facilitated by free-form conversation and makes evaluation difficult. The question then becomes: “is there a better behaviour test than having a free-form conversation?” (Levesque 2014).

In response to this question, Levesque (2014) offers a constructive suggestion. He asks readers to consider the advantages of requiring a machine to answer a directed yes/no question such as, “Can a crocodile run a steeplechase?”—a question to which a person of normal intelligence will answer “no”, by thinking it through.

He continues, “The intent here is clear. The question can be answered by thinking it through: a crocodile has short legs; the hedges in a steeplechase would be too tall for the crocodile to jump over; so no, a crocodile cannot run a steeplechase.” Answers to questions like this will not be available via Google or some other source; rather, the machine has to reason about the physical properties of hedges, steeplechases, crocodile anatomy, and so on. Moreover, the machine must give a yes/no answer; it is not possible to use ‘deception’ and ‘trickery’ to evade the question.

However, questions like this are sometimes amenable to trickery of another sort, “cheap tricks” (aka heuristics). For example, the crocodile question can be answered using the closed world assumption, which says (among other things) the following: “If you can find no evidence for the existence of something, assume that it does not exist.” A cheap trick like this “gets the answer right, but for dubious reasons. It would produce the wrong answer for a question about gazelles, for example.” Accordingly, Levesque proposes a more sophisticated binary-question examination called the Winograd schema challenge (WSC).

## 5 The Winograd schema challenge

A Winograd schema is an anaphor disambiguation problem, consisting of a statement and a question. Each schema comes in two versions, distinguished by a single special word and a correct answer. Levesque (2017) notes that “It is this one-word difference that helps guard against using the cheapest of tricks on them.”

The examples given below appear in Levesque (2011). More Winograd Schemas can be found online (Davis et al. 2019). In the examples below, the alternative special word follows the first in parenthesis; the corresponding answers are designated ‘0’ and ‘1’.

The trophy would not fit in the brown suitcase because it was too big (small). What was too big (small)?

0: the trophy

1: the suitcase

The town councilors refused to give the angry demonstrators a permit because they feared (advocated) violence. Who feared (advocated) violence?

0: the town councilors

1: the angry demonstrators

The lawyer asked the witness a question, but he was reluctant to repeat (answer) it. Who was reluctant?

0: the lawyer.

1: the witness.

Like Levesque’s unusual ‘crocodile’ question, the answers to these questions cannot be googled or scraped and the special words function as a barrier to the use of ‘cheap tricks’. Moreover, the computer must choose one answer or the other; the distractions occasioned by the open-endedness of the Turing test are not possible in the WSC. The computer must answer the questions and correct answers require that the machine think it through. The WSC requires the computer to mimic the ‘humanness’ of everyday reasoning while eliminating the need for the machine to engage in the deception involved in pretending to be human in ways that are not required to demonstrate intelligence. All this renders the WSC, as compared to the Turing test, “less subject to abuse” (Levesque 2011, 2014, 2017).

## 6 Building a working understanding through conversation

In the preceding, we referenced the idea of a practical working understanding between participants in a conversation; however, we have not defined precisely what this means.

Consider this variation of the introductory email exchange between Mike and Susan. It assumes the same backstory as before, except that we replace Mike with an intermediary, Eleanor (Mike’s administrative assistant), who is not known to Susan.

Dear Mike:

Anne asked me to tell the neighbours that Peter died after a struggle with cancer.

Susan

Mike is taking a much-needed vacation, and Eleanor, who handles his email, has been instructed to interrupt him only with important matters. Eleanor replies as follows.

Dear Susan,

This is Eleanor, I am answering Mike’s email while he’s away. He had planned to take a few extra days

and get off the grid, but I can contact him if you wish.  
Has a date been set for a funeral or celebration of life?  
Eleanor

Eleanor understands that Mike’s neighbor has died but does not indicate whether she knows if Mike knew Anne and Peter well: she says, “I can” instead of “I will”, which implicitly asks Susan for further direction, as does the request for information about the funeral/celebration of life.

Susan replies:

Dear Eleanor,  
Thanks for your speedy response. Anne and Peter lived across from Mike until two years ago, but I could not say how close they were to him.  
Susan

Susan responds to Eleanor’s implied question with a fact and an implied question of her own about how close they were.

Eleanor replies to Susan’s email as follows:

Dear Susan,  
If they just lived across the street, Mike would be close to them. Send me the dates and I will forward the message. Have a good day.  
Eleanor

Do Susan and Eleanor have a practical working understanding at this point? They are building one, but an ambiguity remains. It may be the case that Eleanor has assumed that Mike would be close to his across-the-street neighbors. Or, it may be that by ‘closeness’ Susan intends intimacy while Eleanor intends neighborliness. (There is another less likely interpretation we discuss later.)

Depending on context, they may have reached a practical working understanding, even with this ambiguity. At least three scenarios (doubtless readers can think of others) might obtain:

*Scenario 1* Susan really doesn’t much care whether the message gets to Mike or not; she has done her duty by informing Eleanor. For her practical purposes, she and Eleanor have a good enough working understanding that she can send Eleanor the dates and sign off.

*Scenario 2* Susan, out of an abundance of caution, prefers that Mike receive this information. Once again, for Susan’s practical purposes, she and Eleanor have a good enough working understanding that Susan can send the dates and Eleanor can go ahead and pass on the message.

*Scenario 3* Susan is keenly aware that Mike needs an informal stress leave and wants Mike to have this information only if Anne and Peter are quite important to Mike and she is willing to take the responsibility for his not being informed. If Susan suspects Eleanor may be for-

warding the letter based on a misapprehension, she might tactfully respond by rephrasing: “Only if you are certain Mike was good friends with Anne and Peter. I know he needs some rest.”

At this juncture in Scenario 3, the dialog could take many different directions. Susan may even instruct Eleanor not to send the letter.

The preceding illustrates one way a practical working understanding might be built: how problems in constructing this working understanding might arise, be resolved, at least well enough to make a decision, and the understanding sustained, and how a misunderstanding might be identified. With a slight perturbation of the backstories, the conversation could proceed differently.

The Turing test depends on this everyday yet complex human experience of building a working understanding through conversation. Turing’s choice of the term interrogator, together with his use of courtroom language and his comparison of a version of the test to a viva voce examination, indicates this process must be focused. Though he cannot and does not define intelligence or mind, Turing suggests that through a sustained process of questioning, the interrogator can recognize another mind, as revealed through dialog consistent with what the interrogator expects of minds.

If bumps occur in the dialog, the interrogator can smooth things out with further conversation and fix the misunderstanding or realize, in the case of a chatbot, that the limits of the bot’s behaviour have been reached.

Thus, the Turing test begins where the WSC ends: The WSC tests the a priori working understandings of witness and interrogator; the Turing test tests the ability of the witness to engage in a cooperative process of developing a practical working understanding. In and through this process, humans detect intelligence in others. It is a process in which chatbots like Eugene Goostman are unable to participate.

## 7 A practical working understanding

What do we mean by a practical working understanding? This requires explaining two ideas. One is the idea of practical certainty—pieces of knowledge we consider to be certain for practical purposes, since there is little that we, as individuals with limited experience, can verify with absolute certainty.

The other is the idea of a working understanding. What does it mean for two or more people to have a working understanding of an utterance, for example?

We need to understand practical certainty first. We have taken this idea from Henry E. Kyburg Jr.’s (1974) theory of epistemological probability, a variation of which is described by Fahiem Bacchus (1990) for an AI audience.

The idea is dead simple: we accept, as a practical certainty, any sentence that, to the best of our knowledge, has a probability that exceeds some threshold of belief, say 0.95. We are practically certain that we will be able to get milk at the convenience store on the way home, that the subway is running, that our credit card can be used to purchase a meal in Toronto, and that that meal has not been poisoned. We go about our lives as if these things are true, for all practical purposes. (Moreover, we also attribute similar sets of beliefs to other people.) This is not to say that accepting a sentence as practically certain guarantees success (it does not!), but it provides an account of how we navigate efficiently through our corner of a complex universe—we wake up, iron our best clothes, hop on the subway, go to our favorite brunch spot, order and eat a meal, and live to pay for it by taking out our wallet and giving the server a credit card.

But the unexpected happens, often. We oversleep, so we wear whatever is handy. Or the subway station has flooded; we take a streetcar. The restaurant may have changed ownership and closes Tuesdays, so we go across the street. We do not discover until the end of the meal that the new establishment is a small cash-only family business; someone trudges to the nearest automated teller to get some cash. But we had a nice meal, as planned, and will have another one soon. Not everything bad happens on every outing, but the possibilities are legion.

In different settings, we use different thresholds. We will use a different threshold making a medical diagnosis or designing a nuclear reactor than going for brunch. And we have to be aware that no matter how high our threshold of practical certainty, there will be occasions where, if we accept a large number of such sentences, their conjunction will not be practically certain, and, as the brunch example illustrates, something almost always goes wrong in a multi-step plan, but we generally have simple backup strategies.

This is just a thimbleful of Kyburg's theory but gives the general strategy. It plays out in our earlier dialogs, but first, let us clarify what we intend by practical working understanding. We cannot say that one person has unmediated knowledge of what another person intends by an utterance; each builds an understanding of the meaning of the other's communications based on accumulated practical certainties. If the pair has just begun talking, each individual will begin to form practical certainties about what the other intends, and each individual will assume the other is doing the same thing. As the discussion proceeds, each individual's model of the other may become more certain or less, or even qualitatively change, based on their interactions. At a certain point, each individual holds as practically certain a set of sentences about the other, and this is what we mean by a practical *working* understanding. Each party is sufficiently certain that the other's understanding, though not identical to their own, is

sufficiently similar that discussion can proceed to the next step. Still, just like the multi-step dinner plan above, their understandings may diverge at some point and require clarification.

In the first Mike–Susan dialog, the practical working understandings overlap closely, and might be characterized thusly: Mike does not know Anne and Peter all that well, but he may put a card in Susan's mailbox to pass on to Anne, which, in turn, is based on a practical working understanding of social norms.

In the second dialog, Susan and Eleanor build a practical working understanding about whether Mike's vacation should be interrupted. In Scenarios 1 and 2, Susan and Eleanor do not make explicit to each other their practical working understandings, which in all practicality they cannot "see", and which may differ, but both understand, by knowing the desired outcomes obtain, that they are sufficiently similar. For example, suppose Eleanor is a digital assistant, and when she says, "If they just lived across the street, Mike would be close to them," it is a total blunder—she has confused closeness in the sense of intimacy with closeness in the sense of proximity. Their individual understandings of what each intends are, in this case, very different but the desired outcome is still achieved. (In Scenario 3, further work is required.)

Kyburg's theory quickly gets nuanced, and we leave it to the reader to review his original work together with its later modifications, to judge whether it passes muster as a knowledge representation formalism suitable for representing practical working understandings of sentences of natural language. We present his theory, not as definitive (one of us is practically certain it is the right approach and the other is not sure), but as a mature theory of how intelligent agents might build a practical working understanding in an environment where every utterance has ambiguities.

There are different accounts of belief revision, many logic-based, with the AGM model first appearing in Alchourron et al. (1985). Gärdenfors later produced a book (1992), providing an overview of belief revision. Subsequent treatments by de Kleer (1986) and Poole et al. (1987) incorporate a notion of assumptions—possibilities that add to categorical knowledge, and Huang et al. (1991) apply this to user modeling. The probabilistic approach introduces a more permissive notion of inconsistency by allowing generalizations with both subclass and individual exceptions, but that is a topic for another paper.

Interestingly, the idea—though not the terminology—of practical certainty pervades the WSC. Browsing the list of Winograd Schemas (Davis et al. 2019, n.d.), one notices that the 'correct' answers are almost always practically certain. That is, although there is a conclusion the reader naturally jumps to, it is often possible to contrive an understanding that suggests the 'wrong' answer.

Levesque (2011) states that a Schema should be “easily disambiguated by the human reader. Ideally, this should be so easy that the reader does not even notice that there is an ambiguity ...”. He classifies certain schemas as “too obvious”. For instance:

The women stopped taking the pills because they were pregnant (carcinogenic). Which individuals were pregnant (carcinogenic)?

0: the women

1: the pills

Pills cannot get pregnant, and women cannot be carcinogenic. There is nothing to disambiguate. We would say there is no uncertainty; the authors, citing linguistics research, say this can be solved using “selectional restrictions alone”.

He classifies others as “not obvious enough”:

Frank was jealous (pleased) when Bill said that he was the winner of the competition. Who was the winner?

0: Bill

1: Frank

We would say the referent here is, with practical certainty, Bill for the first word (‘jealous’), but just uncertain for the second word (‘pleased’). Levesque makes a similar observation and solves the problem by adjusting the question so that both special words yield practical certainties.

Where Winograd schemas must be constructed by designers so that it is practically certain that each schema contains a pair of practical certainties, the Turing test suggests the actual participants disambiguate uncertainty in real time as conversation unfolds:

Mike cut an opening in the wall for the new window, but it was too big.

A simple question resolves the ambiguity here, but one can imagine many simple exchanges wherein the ambiguity requires some dialog to resolve.

## 8 Summary and conclusions

Levesque (2017) writes that “an informal conversation as suggested by Turing gives a trickster a lot of room to maneuver.” We do not dispute this. And the state of the art at this writing is that it is easier to build a trickster program that steers away from unknown territory than one that can engage in open-ended discourse.

But this is not what Turing intended. Although he suggests the players in the imitation game must use deception to hide the fact that the man is not a woman or that the machine is not a human, this is worlds apart from a program that uses deception to conceal the fact that it is not intelligent. The lack of success thus far only shows that it is still too

soon to expect a machine to pass the Turing test. That said, a machine that could consistently solve WSCs by thinking the answers through would undoubtedly be a remarkable technical achievement; even so, the WSC only tests whether the chatbot has a priori a practical working understanding consistent with that of the test designer; whereas, the Turing test tests whether the chatbot can collaboratively build such an understanding.

Of course, in an open-ended Turing test, Winograd schemas, chatbot dialog, and questions testing a bot’s ability to build practical working understandings are all available to an interrogator, as are other approaches that are certain to arise. This assists the overall goal by giving the interrogator a toolbox of approaches, with the goal of improving the difficulty of the test and the accuracy of the interrogators’ decisions.

What remains remarkable is Turing’s vision—that he saw in a theoretical machine consisting of just a tape and a read–write head that could read, write, and erase binary digits the possibility of a machine that could communicate with humans in meaningful ways, even if today’s talking machines remain wide of the mark—and that he anticipated it would be a hundred years, at least, before machines could stand a chance of passing as human in open-ended conversation.

Just as insightfully, he saw cooperative conversation as a way that we humans ordinarily discern intelligence. The “Mr. Pickwick” fragment depicts sophisticated and comical word play around differing interpretations of “winter’s day” that are finally resolved with the witness stating, “By a winter’s day, one means a typical winter’s day, rather than a special one like Christmas.” Turing’s idea of the interrogator showed that he was interested in more than a talking machine—he wanted the machine to demonstrate engagement with the interrogator. We have characterized the process by which intelligent agents have engaged conversation as building a working understanding.

Some final remarks. By way of reflecting the appearance of voice assistants such as Siri and Alexa in everyday life (at this writing), we crafted the Mike/Susan/Eleanor dialog with the idea of executive assistants in mind—the kind that might eventually be replaced by software agents. Generally speaking, Siri and Alexa do a remarkable job of accommodating a user’s goal provided the goal is within a limited range of tasks that the device understands—calendars, phone calls, text messages, opening apps, and so on, but it does not take much ambiguity for the experience to become as maddening as a menu-based telephone auto-receptionist. The Mike/Susan/Eleanor dialog brings in a set of meaningful social conventions that might vary widely and could not be googled or scraped.

We set out to defend the Turing test against a particular line of criticism. To that end, our focus has been the achievement of a practical working understanding through



conversation, with particular attention to the resolution of ambiguities. However, the relational and dialogic character of human knowing (and human life) has a richness and complexity beyond what we have described. See, for example, Trausan-Matu (2019), Shotter (2019), and Luger and Chakrabarti (2017).

**Acknowledgements** Thanks to David Mould for reminding us of the role of the interrogator and to Wlodek Zadrozny for suggesting the idea of language as collaborative planning. Thanks to the University of Saskatchewan for funding this research, and thanks to the numerous reviewers and commentators on earlier presentations of this work.

## References

- Aaronson S (2014) My conversation with “Eugene Goostman,” the chatbot that’s all over the news for allegedly passing the Turing test. <https://www.scottaaronson.com/blog/?p=1858>. Accessed 15 Jan 2019
- Alchourrón CE, Gärdenfors P, Makinson D (1985) On the logic of theory change: partial meet contraction and revision functions. *J Symb Log* 50:510–530
- Bacchus F (1990) Representing and reasoning with probabilistic knowledge: a logical approach to probabilities. MIT Press, Cambridge
- Commonsense reasoning (2019) Winograd schema challenge. <https://www.commonsensereasoning.org/winograd.html>. Accessed 9 Mar 2019
- Copeland J, Proudfoot D (2008) Turing’s test: a philosophical and historical guide. In: Epstein R, Roberts G, Beber G (eds) *Parsing the Turing test: philosophical and methodological issues in the quest for the thinking computer*. Springer, Netherlands, pp 119–138
- Davis E, Morgenstern L, Ortiz C (2019) The Winograd schema challenge. <https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html>. Accessed 9 Mar 2019
- de Kleer J (1986) An assumption-based TMS. *Artif Intell* 28(2):127–162
- Dennett DC (2012) Turing’s gradualist vision: making minds from proto-minds Turing in context II, Brussels Invited talk
- Gärdenfors P (1992) *Belief revision: Cambridge tracts in theoretical computer science*. Cambridge University Press, Cambridge
- Harkaway N (2018) Will computers be able to think? Five books to help us understand AI. *The Guardian*. <https://www.theguardian.com>. Accessed 12 Mar 2019
- Huang X, McCalla GI, Greer JE, Neufeld E (1991) Revising deductive knowledge and stereotypical knowledge in a student model. *User Model User-Adap Inter* 1(1):87–115
- Kyburg HE Jr (1974) *The logical foundations of statistical inference*, vol 65. Springer Science and Business Media, New York
- Levesque HJ (2011) The Winograd schema challenge. In: *Logical formalizations of commonsense reasoning: papers from the 2011 AAAI Spring Symposium*. Technical Report SS-11-06. AAAI Press, Palo Alto
- Levesque HJ (2014) On our best behaviour. *Artif Intell* 212:27–35
- Levesque HJ (2017) *Common sense, the Turing test, and the quest for real AI: reflections on natural and artificial intelligence*. MIT Press, Cambridge
- Leibniz G (1996) *New essays on human understanding*. In: Remnant P, Bennett J (eds & trans) *Cambridge texts in the history of philosophy*, 2nd edn. Cambridge University Press (**Original work published 1765**)
- Leviathan Y, Matias Y (2018) Google duplex: an AI system for accomplishing real-world tasks over the phone. *Google AI Blog*. <https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html>. Accessed 11 Mar 2019
- Lopatto E (2014) The AI that wasn’t: why ‘Eugene Goostman’ didn’t pass the Turing test. *Daily beast*. <https://www.thedailybeast.com>. Accessed 12 Mar 2019
- Luger GF, Chakrabarti C (2017) From Alan Turing to modern AI: practical solutions and an implicit epistemic stance. *AI Soc* 32(3):321–338
- Marcus G (2014) Why Can’t my computer understand me? *The New Yorker*. <https://newyorker.com>. Accessed 12 Mar 2019
- Martins JP, Shapiro SC (1988) A model for belief revision. *Artif Intell* 35(1):25–79
- Moor JH (2001) The status and future of the Turing test. *Mind Mach* 11(1):77–93
- Piccinini G (2000) Turing’s rules for the imitation game. *Mind Mach* 10(4):573–582
- Poole D, Goebel R, Aleliunas R (1987) Theorist: A logical reasoning system for defaults and diagnosis. In: Cerccone N, McCalla G (eds) *The knowledge frontier*. Springer, New York, pp 331–352
- Shieber SM (2004) The Turing test’s evidentiary value. In: Shieber SM (ed) *The Turing test: verbal behavior as the hallmark of intelligence*. MIT Press, Cambridge, pp 293–295
- Shotter J (2019) Why being dialogical must come before being logical: the need for a hermeneutical–dialogical approach to robotic activities. *AI Soc* 34(1):29–35
- The Society for the Study of Artificial Intelligence and Simulation of Behaviour (2019) Loebner Prize (n.d.) <https://www.aisb.org.uk/events/loebner-prize>. Accessed 12 Mar 2019
- Trausan-Matu S (2019) Is it possible to grow an I-Thou relation with an artificial agent? A dialogistic perspective. *AI Soc* 34(1):9–17
- Turing AM (1950) Computing machinery and intelligence. *Mind Lix* 236:433–460
- Turing AM, Braithwaite R, Jefferson G, Newman M (2004) Can automatic calculating machines be said to think? In: Copeland BJ (ed) *The essential Turing*. Clarendon, Oxford, pp 487–506
- University of Reading (2014) Turing test success marks milestone in computing history. <https://www.reading.ac.uk/news-and-events/releases/PR583836.aspx>. Accessed 12 Mar 2019
- Warwick K, Shah H (2016) The importance of a human viewpoint on computer natural language capabilities: a Turing test perspective. *AI Soc* 31(2):207–221

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