Minds and Machines (2020) 30:637–657
https://doi.org/10.1007/s11023-020-09544-5

GENERAL ARTICLE

Imitation Game: Threshold or Watershed?

Eric Neufeld[1] · **Sonje Finnestad**[1]

Received: 29 April 2020 / Accepted: 1 October 2020 / Published online: 21 October 2020

Abstract
Showing remarkable insight into the relationship between language and thought, Alan Turing in 1950 proposed the Imitation Game as a proxy for the question "Can machines think?" and its meaning and practicality have been debated hotly ever since. The Imitation Game has come under criticism within the Computer Science and Artificial Intelligence communities with leading scientists proposing alternatives, revisions, or even that the Game be abandoned entirely. Yet Turing's imagined conversational fragments between human and machine are rich with complex instances of inference of implied information, reasoning from generalizations, and meta-reasoning, challenges AI practitioners have wrestled with since at least 1980 and continue to study. We argue that the very fact the Imitation Game is so difficult may be the very reason it shouldn't be changed or abandoned. The semi-decidability of the game at this point hints at the possibility of a hard limit to the powers of technology.

Keywords Turing test · Artificial intelligence · Philosophy of mind · Linguistics

1 Introduction

Turing, in a seminal paper (Turing 1950), constructed sample conversational fragments between an interrogator (a rigorous human) and a witness (possibly a digital computer) illustrative of his expectations of machine conversation that might be regarded as human by other humans in conversation with the machine. Here's a snippet:

> *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?*

✉ Eric Neufeld
eric.neufeld@usask.ca

Sonje Finnestad
sonje.f@usask.ca

[1] Department of Computer Science, University of Saskatchewan, 110 Science Place, Saskatoon S7N 5C9, Canada

Witness: It wouldn't 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 don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

Turing considered this an example of a machine giving "satisfactory and sustained" responses. He argued that answering the question "Can machines think?" might require definitions of the terms 'machine' and 'think', which to this day have eluded agreed and satisfactory definitions, and suggested instead the well-known Imitation Game, where an interrogator decides whether an unseen witness he is communicating with is human or not. Such outcomes could be judged statistically.

Some say that a machine *passes the Turing Test* if (say) 30% of judges consider the machine human after a short period of communication, and several competitions described as 'Turing Test' competitions use criteria like this for judging the winner, possibly because Turing predicted that by the year 2000 the average interrogator would not have more than a 70 per cent chance of guessing correctly after five minutes of questioning. To avoid confusion, here we use only the term *Imitation Game*, by which we intend a rigorous interrogator judging a machine human after a sustained engaged conversation, except when citing other work.

Objections to Turing are so numerous and diverse that here we can only provide a sample. We begin with Turing himself, who, anticipating challenges to his audacious claim, asked, "[m]ay not machines carry out something which ought to be described as thinking but which is very different from what a man does?" This is familiar to Artificial Intelligence researchers in the following form: Can a plane imitate a bird so that an observer could not tell the difference? Is that a reasonable test of whether a plane flies? Clearly, no and no. Planes fly but not the way birds fly (French 1990).

Turing concedes that "[t]his objection is a very strong one" but adds, "at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection." In other words, the game is a sufficient but not a necessary test.

Shannon and McCarthy disagree with Turing in their preface to *Automata Studies* (1956) where they adumbrate an objection that claims that the test is not even sufficient. They remark that "[a] disadvantage of the Turing definition of thinking is that it is possible, in principle, to design a machine with a complete set of arbitrarily chosen responses to all possible input stimuli. … With a suitable dictionary such a machine would surely satisfy Turing's definition but does not reflect our usual intuitive concept of thinking."

Gunderson's 'toe-stepping machine' (1964) and Block's "Aunt Bertha machine" (1981) belong to this family of objections, but Searle's Chinese Room argument has enjoyed particular prominence. His thought experiment (or 'intuition pump' (Dennett 1991)) is as follows: Searle, who knows no Chinese, written or spoken, is in

a locked room. He receives pieces of paper on which are written Chinese characters, and returns pieces of paper on which he has written Chinese characters. To an observer, Searle's responses are indistinguishable from those of a native Chinese speaker. Yet Searle has no understanding of the Chinese characters: He only follows a set of rules that tell him what characters to output in response to the inputs he has received; "he has no way to get from the manipulation of the symbols to the meaning of the symbols, and if the person in the room does not understand the meanings of the symbols on the basis of implementing the program, then neither does any other computer solely on that basis, because no computer, just in virtue of its computational properties, has anything that the man does not have" (1990). The same, he adds, is true of the room—or the system—as a whole.

These issues of understanding 'meaning' or symbol-grounding have been raised both by (French 1990) and Harnad (2001).

French accepts the sufficiency of Turing's test but argues that a machine could never pass it because human intelligence as he envisions it has underlying subcognitive and physical levels, the former learned over a lifetime by interacting with the external world in a given cultural context. But the subcognitive is so intertwined with the cognitive and physical levels that the Imitation Game "could be passed only by things that have experienced the world as we have experienced it."

French has revised his assessment since the availability of 'big data' (French 2012), but similar ideas remain extant. E*mbodied intelligence* (Cangelosi et al. 2015) goes back to (Brooks 1990) and argues that a brain exists in a body that interacts with the world, reminiscent also of early notions of the eye-brain connection in computer image understanding (Horn 1985).

Harnad (2001) citing (French 1990) describes the Turing test as a test of 'total indistinguishability', non-verbal as well as verbal. Thinkers, Harnad points out, "can and do more than just talk" (2000). Moreover, our words (or symbols) must be "grounded directly and autonomously in causal interactions with the objects, events and states that they are about (2001)," implying any successful Imitation Game candidate must be embodied, and able to act in the physical world, for example, by reading facial expressions.

In June of 2014, it was reported that a professor at the University of Reading announced that a chatbot known as Eugene Goostman had passed the Turing Test at a competition held at the University of Reading on the anniversary of Turing's death, having convinced 33% of the judges it was human (The Guardian 2014). The same chatbot had finished second at two Loebner Prize Turing Test competitions in 2005 and 2008 and won first prize at another competition at Bletchley Park. The 2014 announcement was initially met with both excitement and scepticism in the mainstream media. However, it appears that Warwick interpreted Turing's turn-of-the-century prediction as doing well at the Imitation Game.

We leave it to the reader to come to an independent decision about such competitions. However, Hector Levesque, a prominent Knowledge Representation scholar, wrote in a series of peer-reviewed articles and a book (2011, 2014, 2017) to the effect that "fooling judges", as exemplified by such competitions, incentivized trickery and deception, and proposed an alternate test designed to keep the computer on

topic. This also caught the attention of the mainstream media (Marcus 2013), and we discuss it in some detail later.

These arguments are difficult, if not impossible, to refute. If we could compile a database with every possible response to every possible statement, we could of course 'in principle' build a machine that could succeed at the Imitation Game with simple lookups. Or could we? Turing expects the machine to be able to sustain a discussion, not just answer questions. So wouldn't we need to be able to store all possible *sequences* of sentences? Is *that* possible?

The replies of the witness in the conversational fragment reproduced earlier are indeed skillful. Not only do they sustain the discussion of an idea, they also demonstrate at least three nontrivial patterns of reasoning or meta-reasoning that transcend ordinary logical deduction yet are commonplace in ordinary discourse, and each of which became a subfield of Knowledge Representation.

That is where we begin our discussion of the game. First we unpack the dialogue quoted above to illustrate (1) reasoning with implied communication, (2) reasoning with generalizations with exceptions, and (3) meta-reasoning, and show that these patterns are not elite tools of philosophers, but occur in everyday reasoning.

We do this in part by placing the Imitation Game in a modern setting—that of an administrative assistant who manages a calendar. We illustrate that the dialog uses the much-studied inference patterns just noted. Afterwards, we consider three specific critiques of the Imitation Game from long-time practitioners within the AI community. Although they give different perspectives, it is interesting that the Imitation Game has confounded computer scientists dedicated to AI, and striking in light of the progress in areas like networks, computer graphics, and media. Is it possible that the simple ability to engage in communication with another living being represents a hurdle that our present-day notions of automata might approach but never clear?

Posing this question does not obviate any research direction currently underway, no more than understanding the complexity classes P and NP has deterred computer scientists from tackling hard problems. Rather, it has guided them, and the same could be said of decidability. Answering Turing's challenge will offer deep insights into strengths and limits of both humans and machines.

2 Inference Patterns Used by Turing's Witness

2.1 Implied Communication

Besides owning explicit expertise in domains ranging from relatively sophisticated information about literature to mundane knowledge that Christmas as a special day in the otherwise bleak winter, Turing's witness and interrogator both make interesting inferences about implied but unspoken communication. For example, the interrogator's opening sentence is a straightforward yes/no question. The witness doesn't answer it, but heads directly to the explanation "It wouldn't scan", which implies that the answer to the asked question is "no". In constructing an answer, the witness makes the inductive inference that the interrogator will make the inductive inference from the witness's explanation that the witness's implied answer is "no".

A classic example of this pattern sometimes used in introductory AI classes is a machine answering "Here you are" to the request "Can you pass the salt?" rather than "yes". The following fragment (based on (Allen and Perrault 1980)) contains many instances of implied communication:

Customer (breathless): 3:15 to Windsor?
Clerk: Three dollars. Gate 10. Run!

The reader likely pictures a train station immediately.

How would one tackle the problem of designing software to respond like this clerk? First, the computer would need meticulous details about the ticket wicket setting, a "straightforward but tedious" task (as mathematicians sometimes say of lengthy proofs), which one senses has begun already with products like Siri and Alexa. Parsing the utterance "3:15 to Windsor" is more difficult. The clerk must determine that the customer wants to purchase a ticket for a train before it leaves and that "3:15" is not a price or a train number. The customer is not likely asking *whether* there is a train leaving for Windsor at 3:15. If the discussion allows completely open-ended statements, the possibilities become endless: perhaps the customer should be at the bus station, not the train station, or to give a completely off-the-wall example, is giving the name of a song or movie (*3:10 to Yuma* was both*)*.

Allen and Perrault attack the problem using an approach similar to collaborative planning, whereby participants infer goals from statements, then formulate a plan to overcome obstacles (e.g., not having a ticket). A previous paper (Neufeld and Finnestad 2020), discusses language as collaborative planning in the context of the Imitation Game, and using Kyburg's (1974) epistemological probability as a conceptual framework to understand how participants jointly build a practical working understanding of a common goal.

3 Reasoning from Generalizations with Exceptions

Towards the end of the introductory dialogue fragment, the witness employs a second inference pattern—reasoning from generalizations with exceptions that may themselves have exceptions: a pattern also ubiquitous in ordinary conversation.

We presume the witness and the interrogator share the following background knowledge:

Winter days are bleak.
Christmas is a winter's day.
Christmas is not bleak.

"Winter days are bleak" is a generalization with exceptions (such as Christmas). "Christmas is a winter's day" is a fact because Christmas occurs during the winter season. The last sentence, like the first, might also be a generalization with exceptions: many of us have experienced a bleak Christmas.

During the 1980s, this reasoning pattern became known within the Artificial Intelligence community variously as nonmonotonic logic (McDermott and Doyle 1980), default reasoning (Reiter 1980), defeasible reasoning (Loui 1987), common sense reasoning (McCarthy 1986), and specificity (Poole 1985). Anecdotally, we remark that it dominated research within the Knowledge Representation (KR) community for some time. Because these formalisms all have slightly different definitions, for simplicity of presentation we refer to this pattern as *reasoning from generalizations with exceptions*.

Writers of that era noted that this reasoning pattern is easily learned and understood by humans, but designing a software tool to manage knowledge in this form proved to be a challenge. The reader likely can see that converting the sentences to an obvious first-order logical form immediately creates an inconsistency. (Christmas day is not bleak, because it is Christmas, yet Christmas is also bleak because Christmas is a winter's day and winter's days are bleak.) The KR community persisted in a search for a quasi-logical solution for some time, perhaps because numerical probabilities had been dismissed as "inadequate for artificial intelligence" (McCarthy and Hayes 1969), or perhaps because ordinary persons understand the intended meaning this inference pattern without needing to collect statistics.

The initial formalisms proposed by the KR community were received enthusiastically by the AI community after performing well on small examples, but in due course side effects revealed themselves. Attempts to eliminate them resulted in a regression whereby the repairs just pushed new side effects out elsewhere. We present a detailed example of this in Appendix 1.

Kyburg's epistemological probability (1983) treats generalizations as statistical statements. He proposes that sentences of probability (say) 0.95 are *practical certainties*, (certain for practical purposes) and posits that two practical certainties cannot be conjoined unless the resulting conjunction is also a practical certainty. His work did not suffer from the technical problems mentioned earlier (see Appendix 2, which explicates Kyburg's ideas in some detail.) The Imitation Game then can be understood as follows: witness and interrogator build practical working understandings of a problem and a goal. Each participant may have a slightly different understanding but where this creates problems, the parties can then better align their understandings through discussion. Even then, alignment need not be perfect, just sufficient to move the task forward (Neufeld and Finnestad 2020).

The discussion between the interrogator and the witness about winter's days and Christmas provides an example of how generalizations with exceptions are handled in ordinary dialogue.

The way this knowledge would typically play out in the discussion of a KR formalism is that an agent would have background knowledge of two practical certainties: the generalization about winter's days ("Winter's days are bleak"), and the generalization about Christmases, which is an exception to the previous generalization ("Christmas is not bleak"). It would also contain the categorical fact that "Christmas is a winter's day". After constructing the inference machinery, a question and answer scenario would be expected to play out as follows. The answer to the question "Is a

winter's day bleak?" should be "yes", as would be the answer to the question "Is a December day bleak?" The answer to "Is December 19 bleak?" would also be "yes", since December 19 is a winter's day and the exception cannot be applied in these cases because the antecedent (*Christmas*) is unknown or known to be false. The question "Is Christmas bleak?" could be answered *yes* or *no*, because either practical certainty could be applied. However, it is generally agreed that given two practical certainties to choose from, one should choose the one based on the narrowest class. So, the answer to this query is "no". This can continue through many levels. If it is also known that Christmas 1946 was bleak, the answer to the previous question will still be *no,* because the rule about Christmas is the narrowest that can be applied. But the answer to *"Was Christmas 1946 bleak?"* will be "yes", not because that day was a winter's day, but because in this case, "Christmas 1946 was bleak" is the narrowest statement that can be applied.

Kyburg justifies the rule about choosing the generalization based on the narrowest class by introducing a notion of epistemological randomness. Given the two practical certainties above, one can assume that if a winter's day has been selected, it has been randomly selected if there is no information to contradict that assumption of randomness. If the day selected is known to be Christmas, then the broader generalization cannot be used because it can no longer be assumed to be random. However, it *can* be assumed to be a typical Christmas day. We remark that Kyburg's treatment has other nuances that we do not address here, some of which appear in Appendix 2.

We now show that in the present witness/interrogator interaction, the use of generalizations with exceptions plays out in ways far more sophisticated than the question/answer format just described.

4 Reasoning About Reasoning

At the end of the dialogue fragment, the witness goes beyond reasoning with domain knowledge and/or reasoning from generalizations with exceptions to challenge the reasoning style of the interrogator ("I don't think you're serious"), and then explains the witness's thinking ("By a winter's day one means a *typical* winter's day, rather than a *special* one like Christmas." (Emphasis ours.)).

This is called meta-reasoning. The parties might be thinking as follows: The witness believes the interrogator is baiting him or testing him by presenting a possibly foolish or playful argument, so the witness says so ("I don't think you're serious"). The witness believes that the interrogator is trying to argue that because winter's days are bleak, and Christmas is a winter's day, and Christmas is special, a winter's day *can* be special and being compared to a winter's day need not be objectionable. This is true, but then the witness clarifies what is meant by *"a winter's day"*—namely a typical (or randomly selected) winter's day, in effect instantiating the rule for reasoning about hierarchies of exceptions.

There are other plausible ways to model this conversation. In another context, the interrogator might not be aware that Christmas is not a typical winter's day. In

that setting, the witness might drop the sharp remark, "I don't think you're serious", and, in the process of building a better practical working understanding, just gently remind the interrogator about the nature of generalizations and the exceptional aspect of Christmas. This also reinforces the collaborative nature of speech: as an issue arises, the parties work to align their working understandings.

Meta-reasoning arises in AI treatments of user modelling (van Arragon, 1991). Tutoring software, for example, might try to construct a practical working understanding of the possible mental model of a student who has given an incorrect answer to a problem. This makes it possible for a tutor to go beyond understanding that the student's answer is wrong and begin a collaborative process that can troubleshoot the student's mistake. However, as Turing's example shows, meta-reasoning arise whenever two agents differ.

5 A "Practical" Application

Turing's conversational fragments may not resonate with engineers looking to create practical products for the marketplace. To address this, we take the Imitation Game from Turing's drawing room to the modern office where the task at hand is maintaining an employer's calendar via email instructions, since digital assistants are presently so popular (but limited). Where Turing's witness focused on the goal of sustaining an academic discussion, our assistant's ultimate goal is scheduling—and keeping clients happy. Anyone who's enjoyed the benefit of a good administrative assistant knows how challenging this is.

Here is dialogue from a setting where the responding party (if a machine) would be considered to be doing well at the Imitation Game. The backstory is that Susan, who lives near Mike, wishes to share with Mike information regarding the death of a former neighbour. Unbeknownst to Susan, Mike has decided to get off the grid and asked his assistant, Eleanor, to manage his personal email, and has issued the following instruction:

> I don't want to be interrupted while I'm away,
> Unless something important comes up.

This is not an uncommon instruction to give an assistant, and it follows the pattern of a generalization with exceptions, a different analysis to that discussed in (Neufeld and Finnestad 2020). The pursuant commentary on the dialogue was crafted to illustrate the structure and variety of inferences in such an ordinary task. Though such a discussion could go in any number of directions, we illustrate a path whereby the parties arrive at a shared understanding and get to the goal of signing off.

Susan writes Mike:

> *Dear Mike:*
> *Anne asked me to tell the neighbours that Peter died after a struggle with cancer.*
> *Susan*

Eleanor replies:

Dear Susan,
This is Eleanor, I'm answering Mike's email while he's away. He'd 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 may have inferred from the fact that Susan informed Mike that Peter died "after a struggle with cancer" that Mike may have been unaware that Peter had cancer, implying some distance in their relationship. Or this may be her standard response. At this point, it doesn't really matter. Susan knows Mike well enough to understand from Eleanor's response that Mike has handed over his personal emails to Eleanor. From Eleanor's response "I can contact him if you wish", Susan infers Eleanor has chosen not to immediately forward Susan's message to Mike, whatever the reason. Thus implied communication has already taken place. But note that Eleanor delivers the message with an implicit invitation for Susan to present a case for an exception.

Susan replies:

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

Susan responds to Eleanor's implied question with straightforward factual information that Anne and Peter were Mike's neighbours, which could be interpreted as an implied request to Eleanor to use this new information, along with her knowledge of Mike, to make the call as to where Mike is on the continuum between those who would get to know their neighbours and those who would not, and act on it.
In the answer below, Eleanor (perhaps) judges that the gregarious Mike would know his neighbour well, enough to trigger an exception, and now explains her reasoning:

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

If Eleanor were a digital assistant playing the Imitation Game, her performance would compare favourably with that of Turing's witness, and would be qualitatively superior to current consumer level digital assistants. Moreover, Eleanor uses reasoning patterns similar to those of the witness. (Note: the astute reader might have noticed the possibility that digital Eleanor may have goofed by interpreting 'close' as a measure of distance. In this setting, Eleanor gets away with the error; in another, she might not.)

6 Arguments Against the Imitation Game from Computer Science

Despite the possibilities a machine capable of playing the Imitation Game well would offer in practical settings, many arguments have been advanced against the Game within the computer science and artificial intelligence communities. These arguments do not reflect disagreement with Turing's wish to sidestep the tricky matter of defining intelligence. Rather, all three positions presented here raise questions about the practical benefits of pursuing the Game in order to achieve such objectives.

7 Levesque: The Imitation Game is Too Easy

Levesque's (2011, 2014, 2017) objection is that the Game incentivizes *deception*: the computer can succeed by fooling observers into thinking it is human. Levesque references the 'trickery' and 'evasiveness' of the chatbots in the Loebner competition and others, writing that they respond with "elaborate wordplay, puns, jokes, quotations, clever asides, emotional outbursts, points of order. Everything, it would seem, except clear and direct answers to questions" (Levesque 2011).

We have discussed Levesque's argument in detail elsewhere (Neufeld and Finnestad 2016a, b) but sum it up here as follows: In describing the Imitation Game, Turing suggests certain protocols may have to be followed so that a machine doesn't immediately give away its identity, say, by doing a difficult calculation in a fraction of a second. Pausing for a few seconds, or pretending to backspace and correct a typing error, is in keeping with other protocols like communicating by teletype rather than speech and not being visible to the interrogator; the machine uses deception to hide the fact it is a *machine*. But nothing in Turing's work suggests that the machine use deception to disguise the fact it isn't *intelligent*; to the contrary, Turing's fragments demonstrate his high expectations for "satisfactory and sustained" responses.

Levesque's critique is constructive and proposes an interesting alternative to the Imitation Game, the *Winograd Scheme Challenge* (WSC), a exam of cunning multiple-choice questions that forces the machine to concisely answer the questions posed to it. A WSC question has two forms that differ by a single word, and for each question-form, only one of two possible answers is correct.

In the example below, one form of the question omits the word in brackets; the other omits the word preceding it. The answers corresponding to each form of the question follow.

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.

It the trophy would not fit in the suitcase because *it* was too small, *it* must refer to the suitcase; alternately if the trophy would not fit in the suitcase because it was too *big*, *it* must refer to the trophy.

A pronoun should replace a noun only if its referent is unambiguous, but this often requires the reader to understand the full meaning of the sentence. Some referents can be easily guessed using just number ("Mr. Smith did not hire Adair graduates because they hadn't studied analytical geometry") or gender ("Mr. Smith defibrillated Ms. Jones after her dramaturg contracted reflexive sympathetic dystropy"). WSC questions should be easy for a human to answer, but hard enough that a machine needs to understand the sentence's meaning. Getting the correct answer requires 'thinking it through', claims Levesque.

This prevents chatbot-type trickery, but the flipside is that the one-word answers gives little insight to the judge trying to determine whether the machine actually 'thought it through', or can sustain an discussion.

That said, Levesque's constructive proposal has already inspired an international competition, which, though not as entertaining as the Loebner competition, has potential to generate useful milestones towards the Imitation Game.

8 Ford and Hayes et al.: The Imitation Game is Too Hard

Over a period of time, Ken Ford, Patrick Hayes and other co-authors (Ford and Hayes 1995; Ford et al., 2016) have put forward a theory of *cognitive orthoses* (previous called *cognitive prostheses)* as an alternative to the Imitation Game, a program very different from that of Levesque and with different motivation.

These authors argue very strongly against the Game. The Game creates a regression: it avoids the problem of defining 'intelligence' but introduces the problem of defining what it means to "do well". They (1995) also state that the Imitation Game tries to detect that there *isn't* a difference between two behaviours, that is, it tries to confirm the null hypothesis. Additionally, they believe that it is also necessary to 'judge the judge': in judging the witness, one must also consider whether the judge was clever enough to ask sufficiently telling questions. This means the skills of the interrogator must be measured, which means the idea of 'telling questions' must also be formalized. (This could also be an argument against the WSC!).

They do make the interesting observation that at the time Turing was working, the ability to perform mental arithmetic, now considered mechanical, was evidence of intellectual ability, thus introducing a 'shifting sands' argument against the Imitation Game: as AI software improves, so will the ability of judges to detect flaws in the behaviour of nonhuman machines, and consequently, machines may never be judged as doing well at the Game. (This may occur at a subconscious level, analogous to the "uncanny valley" (Mori, 2012) experienced by consumers of computer renderings of humans and virtual reality (Lanier 2017).) Ford and Hayes do not mince their words: the Imitation Game "is not a sensible goal for Artificial Intelligence"; it may be "actively harmful".

They go on to say: "Whether or not he intended it, [Turing's] insight that technology might … reach a kind of divine power almost certainly played a key role in motivating early AI projects." However, they say, the time has come to "distinguish legend from science." Where Levesque felt the Imitation Game was too easy

to game, Ford and Hayes feel the Game aims at something almost godly that will get only harder with the passage of time.

Ford and Hayes fortunately offer a constructive direction in which to take AI instead: the production of cognitive orthoses—devices that amplify our own cognitive abilities, the way eyeglasses improve our vision. This computational goal is "now and near, a computational Golem is not". For example, though services like Alexa, Siri and autocorrect are a far cry from ideal, they are much better than would have been imagined thirty years ago.

We like the metaphor of Ford and Hayes—it is not that different from human-in-the-loop computing, where software and humans work as a team. The machine handles the heavy-duty computation, and the human communicates hunches through an interface as to how to direct the computation, and both tasks are easily managed on modern consumer hardware. For an example, see (Núñez Siri et al. 2020). However, we don't see that this idea obviates the value of the Imitation Game. In fairness to Turing, the arguments from Ford and Hayes about the vagaries within the Game also apply to cognitive orthoses. Who judges, and by what criteria, whether a piece of software has succeeded as a cognitive orthosis or not? At the extremes, a case can be made that every reasonably good piece of software amplifies some cognitive ability, and often, another case that every piece of software stunts some other cognitive ability: software still often forces the user to trade efficiency for the ability to handle special cases. In that case, everything is a cognitive orthosis, or nothing is, or anywhere between.

9 Stuart Russell: The Imitation Game is Terrible

Stuart Russell is well-known in the AI community as the co-author of *Artificial Intelligence: A Modern Approach* (Russell and Norvig 2009), a comprehensive and practical treatment of Artificial Intelligence that is also one of most widely used undergraduate/graduate textbooks. A scholar in his own right, he is Vice Chair of the World Economic Forum's Council of AI and Robotics, and he has worked with the United Nations. These credentials suggest that he has a good understanding of the state of the art in AI at this time and the likely rate of progress.

However, we found Russell's precise position on the Imitation Game difficult to understand. He (2019) characterizes the Imitation Game as "a thought experiment designed to deflect skeptics who supposed machines could not think in the right way, for the right reasons, with the right kind of awareness." We couldn't agree more, but then he plainly says Turing's Imitation Game is "a terrible definition to work with", and add that researchers instead are pursuing the automation of rational behaviour. But then, in an unexpected twist, he says that one of the first major breakthroughs required to achieve this goal is understanding language and common sense: practically a restatement of the Imitation Game.

In any case, Russell is highly optimistic about AI's future, suggesting superintelligent computers may appear within the lifetime of his children, making him just a little more conservative than Turing, who in 1951 estimated completely open-ended conversation would take at least 100 years to achieve. Russell doesn't express

interest in language as an interactive tool for collaborative planning, but he believes it has great value as a way for machines learn quickly. He writes, "A machine that really understands human language would be in a position to quickly acquire vast quantities of human knowledge", as opposed to learning concepts from sensor data. He points to Project Aristo, which aims to build systems that can pass science exams after reading textbooks and study guides. He provides the following test question:

Fourth graders are planning a roller-skate race. Which surface would be the best for this race:

(A) Gravel
(B) Sand
(C) Blacktop
(D) Grass

This resembles Levesque's WSC questions. Once the machine has read its textbooks, it must 'work out' that a roller-skate race is a race between people and not roller skates, and that the surface referred to is the racing surface, not where spectators will sit. It is difficult to see how this process differs from Turing's witness or Eleanor, albeit in other domains, and, as with the WSC, there is no way to tell from the answer whether the machine 'worked it out'. Although the test question is a basic science word problem, Russell's description of the machine's reasoning suggests that the machine needs a broad base of common-sense knowledge to clearly understand the problem before choosing its answer.

Where to from here? Russell's imagination is not limited by Aristo. As one goes through *Human Compatible*, one finds fodder for a thousand novels as Russell plunders the treasure chests of philosophy and psychology from Aristotle to Kahnemann, explaining the potential and perils of future superintelligent machines as they tackle challenges where the stakes are far greater than the parlour conversations of Turing's witness, or the calendar of Eleanor's office. To sketch one scenario, Russell posits that an intelligent machine's first priority must be to accomplish its master's goals. But a self-driving car taking its master to the airport shouldn't have to be reminded to obey laws and drive safely. Perhaps the robot assistant needs a social conscience. Or does it? Might the self-driving car skate as close to the edge of the law as possible, driving with wild abandon if that benefits its master? In fact, might it advance its master's position illegally if it can do so without getting caught? Might this extend to the political sphere?

There is also the thorny problem of decision-making under uncertainty. Will the superintelligent pilot of a self-driving Phantom 309 swerve to miss a full school bus, sacrificing itself, its human copilot and its valuable cargo? Would it allow the human copilot to override its decision? Questions like these are bound to arise in the discussion of self-driving cars.

Where certain AIers have suggested these kinds of possibilities are just around the corner, Russell is candid about the challenges ahead. One repeatedly reads in his work sentences like: "AI researchers are only just beginning to get a handle on how to analyze even the simplest kinds of real decision-making systems, let alone

machines intelligent enough to design their own successors. We have work to do." And again, "the robot butler, managing the household with aplomb and anticipating its master's every wish, is still some way off—it requires something approaching the generality of human-level AI".

Russell implies that these machines will not only be smart, they will be super-smart, and super-fast, so much so that we may need to take steps in advance so that they do not become our masters. Yet we ask: apart from the high stakes they are playing for, are Russell's machines so different from our Eleanor, who would be expected to be good at a relatively straightforward task that requires good interaction skills. Might Russell's machines 'merely' need Eleanor's understanding of the world to bootstrap themselves into a position of authority? Will Eleanor eventually take over the office and then the world?

As we approach the divine, the devil seems to be in the details.

10 Where Does Eleanor Fit In?

Just as Turing's witness must understand whether a particular metaphor will be interpreted as a compliment, Eleanor while dealing with Susan must somehow have prior knowledge that a funeral is a meaningful event in the life of a community. She also should know that it is inevitable in the life of an individual. That is, death is on the one hand a significant event, and on the other, ordinary. That sounds hard-hearted, but Eleanor must be able to gauge how important this event is to Mike at this particular moment.

There are other complications. Is Mike absent because he is on the verge of physical and mental collapse from the rigours of his high-pressure executive position and desperately in need of rest? Or is he just headed to Vegas with friends for a few days? How great a knowledge base do we need to enumerate enough possible states for Mike that an automated version of Eleanor can do reasonably well at calibrating Mike's needs?

If Mike is to be notified of *all* deaths, that solves one problem, but consider all possible events that Susan or another caller might communicate to Eleanor: a cancer diagnosis, a change in the weather, a birthday, a divorce, the loss of a set of keys, the disappearance of a pet, a bargain. We'd wager that a few of these events bring a complete scenario to the reader's mind. How many such scenarios need to be explicitly enumerated and added to the knowledge base and at what level of complexity? Might it be possible to template or generalize them for the purpose of efficiency? If so, how are unanticipated scenarios slotted into the templates?

So far, we have only talked about how Eleanor should gauge Mike. Suppose Mike is given a different portfolio. Would a general-purpose Eleanor have to be able to learn the unique preferences of a new supervisor? Or the new portfolio?

The amount of necessary background information becomes staggering as soon as one leaves the toy examples, and so far, we are just discussing how Eleanor will deal with a single instruction.

Turning to the problem of designing Eleanor's reasoning, we also note that the three patterns of inference discussed earlier—implied information, generalizations with exceptions and meta-reasoning—are not easily disentangled. Let's revisit the point where Turing's witness says, "It wouldn't scan", rather than, "No. It wouldn't scan." Here is a sketch of a possible deep representation of the background knowledge:

The witness, in replying to the interrogator, assumes that the interrogator is using a generalization about meta-reasoning along these lines:

If there is no clear binary answer given to a binary question.
Then infer the answer from the context of the reply.

A generalization like this may have an exception. For example:
If there is a one-word response to a binary question of negative polarity,
Ask the respondent to clarify.

In the case of Eleanor, the exception would help her respond to a superior asking 'Aren't you going to answer Susan?', where a simple 'yes' could mean 'yes, I am', or 'yes, I'm not', depending on whether the respondent is replying with the truth-value of its situation or with the polarity used in the question.

Clearly the inference engine is getting complicated, too.

Earlier we mentioned that Searle (1980) famously poses problem of an agent (himself) who receives queries written in a foreign language he doesn't understand, which he answers by looking up the question in a massive rule book and copying the answer. Is the agent intelligent?

Searle answers *no*, because the agent doesn't understand what is transpiring even though an observer, unaware of the internal workings, would judge the agent's responses as indistinguishable from those of a native speaker of the language.

Levesque (2009) responds "There's no such rule book!" We agree! But we (Neufeld and Finnestad 2016c) suggest that if the agent is perceived to be doing well, the Imitation Game has been solved whoever or whatever wrote the rule book! It's a homunculus problem! In which case, we'd concede that the agent also passes the Imitation Game, even if the agent is unaware of what is being accomplished. However, the point of the Imitation game is precisely to get away from notions like consciousness and awareness.

But then, is Eleanor possible? Is there no such rule book (knowledge base) to guide her modest tasks? The preceding discussion suggests that writing a manageable number of typical scripts, given the variety of human experience, would at the very least be difficult.

Back to Ford and Hayes, if Eleanor's wrong decisions do not have significant impact beyond mildly inconveniencing Mike, it might be possible to build a useful syntactic Eleanor that acts as a cognitive orthosis for Mike that does well often enough, but is unpredictable in unfamiliar situations, as are the present generation of voice assistants and tools like auto-correct. A general-purpose Eleanor that can consistently respond to emails as illustrated, to use Russell's expression, appears to be a long way off. We have to concede that while epistemological probability avoids problems created by using ordinary logic to construct practical working

understandings, a closer comparison of Eleanor and Russell's machines (see (Shotter 2019; Trausan-Matu 2019)) suggests there is much more at play in these examples.

11 Conclusions

The Imitation Game raises many questions; the arguments sometimes seem to go in circles. But we are practically(!) certain of the following.

The game doesn't harm progress. The world's largest corporations have invested heavily to develop computational prowess at chess, Go, Jeopardy, and making hair appointments (Leviathan and Matias 2018), even though there has been little, if any, success at creating an agent as capable as Turing's witness. This fact alone speaks volumes. Clearly, the challenge has inspired much computer science and philosophy.

Turing's Game needs no significant redefinition. Certainly he put forward his ideas in plain language at a different time, and, if pressed, we *might* rephrase it in more contemporary language by saying that a machine can be judged to have done well at the Imitation Game if it consistently judged to be human after engaging in sustained conversation. But it is likely Turing was very careful in his description of the Game. The fundamental model of computation named after him is the cornerstone of computability theory, and the Church-Turing Thesis (Kleene 1952), which also bears his name, states that a function on the natural numbers can be calculated by an effective method if and only if it is computable by a Turing machine. Turing also understood the limits of computing (Turing 1936), by demonstrating impossibility of the halting problem. Nor was he naïve about probability and statistics, having collaborated with I.J. Good at Bletchley Park on what is now called Good-Turing frequency estimation (Good 1953) (which curiously has found its way into recent work in statistical linguistics). Thus, it is practically certain that when Turing in his 1950 paper repeatedly used the phrase "doing well at the Imitation Game", he was carefully articulating a problem description based on an extraordinary background in mathematics and computation. His game is not so different from the idea of judging "the realism of an artificial image", which cannot be measured, yet photorealistic computer graphics thrives as practitioners deepen their understanding of how light interacts with surfaces and how the human perceptual apparatus responds to the renderings produced by computers.

There is no need to 'shut down' any other avenue of investigation. Embodied intelligence, for example, takes on new meaning when one reads of spiders and mice awkwardly adjusting to life in zero gravity on the International Space Station (Kelly 2017; Wired 2011). Consider the size of a spider's embodied brain, then consider the approach to the hungry-monkey/banana problem (Feldman and Sproull 1977) which includes lots of LISP code. Nothing in that approach would help a cyber monkey reach the banana in zero gravity if no chair was available.

It is reasonable to ask whether the Imitation Game defines a watershed, rather than a threshold towards which researchers can consistently expect make incremental progress. Given watersheds in complexity, researchers have successfully found

ways to mitigate the practical and theoretical challenges of provably hard problems. Calude (2018), for example, looks at probabilistic approaches to the halting problem. One might even consider the area of software testing as discovering heuristics to solve the undecidable problem of program correctness. But can they always? While there is no proof that humans are more powerful than Turing machines, there is no proof that they are not.

Whatever the ultimate outcome—whether success is finally achieved, or if a clear reason is identified that explains the difficulty—it will offer an interesting insight into the nature of human thought.

Acknowledgements As with the knowledge we expect Turing machines to master, many of the statements herein may raise questions about exceptions and edge cases. We found it necessary, for example, to not present Kyburg's entire opus, which he spent a lifetime improving. Here, Kyburg's formalism provides a conceptual framework for understanding the nature of the problem, and we have elaborated sufficiently to address the examples presented in the main body. "Phantom 309" is a Red Sovine tune about a ghost truck, the driver of which sacrificed his life to save a bus full of children. The Phantom 309 still haunts the west coast, picking up the occasional hitchhiker and giving him a little change for a coffee. Thanks to Rosemary Nixon for a careful edit, Braden Dubois for several reads and re-reads, the reviewers of this paper for their comments, and thanks to the many persons we have discussed this work with over the years. Thanks also to the University of Saskatchewan for providing funding for this research.

Appendix

Appendix 1: Representing Generalizations with First Order Logic

Summing up the problems encountered during a decade of research necessarily requires oversimplification. We begin with a classic example:

Chilly-Willy is a penguin.
Donald is a duck.
Penguins are birds.
Ducks are birds.
Birds fly.
Penguins don't fly.

We leave it to the reader to observe that by combining different subsets of this knowledge base, we can show that Chilly-Willy flies, and that Chilly-Willy does not fly. We can eliminate one of these conclusions by designating "birds fly" as a generalization, and adding a rule that an instance of a generalization can only be applied if the set of all sentences used cannot be made to generate a contradiction. We can still use an instance of the generalization to conclude Donald is a bird and can fly. These sentences allow other interesting inferences. For example, taking contrapositive forms, we conclude that if it flies, it's not a penguin, and if it's not a bird, it's not a duck. In this setting, the contrapositive forms make sense, but that isn't always the case.

Now suppose, ducks are different from most birds in that they have webbed feet. To fully incorporate this into the database, we add the following:

Chilly-Willy is a penguin.
Donald is a duck.
Penguins are birds.
Ducks are birds.
Birds fly.
Penguins don't fly.
Birds don't have webbed feet.
Ducks have webbed feet.

Again, this seems reasonable. But suppose we add one more sentence.

The only birds are ducks and penguins.

Although it seems unreasonable to say every bird is either a duck or a penguin, this gives a compact counterexample. Let's explore the problem with the compact counterexample, then generalize to something more reasonable.

The counterexample goes as follows. Let Foghorn be a bird. Birds typically fly, so if Foghorn flies, Foghorn can't be a penguin. Because everything is either a penguin or a duck, Foghorn is a duck and has webbed feet. Using the same trick, we can put together a different set of sentences and conclude that Foghorn is a penguin and can't fly.

Here is a fuller counterexample. Let there be 1000 kinds of birds. Each kind is different from a typical bird in some way, but otherwise is a normal bird. (If the kind has no distinguishing feature from other kinds, how can it be a kind?) Plus, we have the clause that every bird must be one of the 1000 kinds. Using 999 of the generalizations contraposed, we can rule out that Foghorn is not any of *those* 999 kinds, and therefore must be the remaining kind and be unique in the way the remaining kind is unique.

Some readers will see this as a variation on Kyburg's lottery paradox, others as a variation on Simpson's paradox. Either way, the simple reasoning pattern initially proposed has collapsed. This result is our interpretation of (Poole, 1989).

Appendix 2: What Practical Certainty Buys Us

The idea of practical certainty lets us hold as beliefs a *set* of sentences, which written as a conjunction would contradict some fact. The classic example is a lottery, where the purchaser buys a numbered ticket, and only one number wins, as opposed to modern lotteries where the purchaser can choose their numbers. It is reasonable in such a lottery to believe, for each ticket, that it will lose. But it is not reasonable to believe that no ticket will win, since by construction a winner is drawn. (For an argument that this still holds for the modern lottery, see (Neufeld and Goodwin 1998)).

To keep the calculations simple, let's suppose there are 20 unique tickets in the lottery. This means that the probability any ticket will lose is 0.95. Thus, it is practically certain that each ticket loses. Suppose an individual buys two tickets. The

probability that both tickets lose is 0.9 – this is *not* a practical certainty, but a probability. In this situation, we cannot combine two practical certainties into a conjunction that is a practical certainty.

As the number of tickets gets large, one can be practically certain that if two tickets are purchased, both will lose.

Applying this to the previous 'bird' example, we can't treat the conjunction that Foghorn is not any one of 999 kinds of birds as a practical certainty. This prevents the collapse of the formalism, but also limits the formalism's inferential power.

We remark that the example above assumed buying tickets without replacement, which simplified the calculation. More generally, let A and B be any two events of probability 0.95. Thus each is a practical certainty. Using the basic identity.

$$P(A\&B) = P(A) + P(B) - P(A \text{ or } B).$$

(where P is probability) the probability of the conjunction could be less than 0.9 because P(A or B) might be 1.

However, if A and B are two arbitrary events of probability 0.99, we can show the lowest value of their conjunction is 0.98 using the same formula, even if the probability of the disjunction is unity, and the conjunction is a practical certainty.

Finally, we remark that the theory of epistemological probability has many nuances. A reader of an earlier draft of this paper asked the following. If 1% of ticks carry Lyme disease, then 99% of ticks do not, and thus it is practically certain that ticks do not carry Lyme disease. This is a knowledge engineering problem worth delving into.

We will use natural language representations of the knowledge rather than introduce a new formalism. To begin with, suppose a data collector has written "Of 300 ticks examined near Gormley Wood, 3 carried Lyme disease". This might be translated to "The probability of any particular tick carrying Lyme disease is between 0.009 and 0.011", the interval accounting for all manner of uncertainty about how the data was collected. Next we learn, "Alice Butterwick noticed a tick on her dog Mollie after a walk through Avon Gorge." From the statistical data, Alice can infer that "the probability the tick on *Mollie* has Lyme disease is about 1%" (this is *lifting*) and therefore be practically certain that *that* tick does not carry the disease. If three hundred dog-owners visit the Gorge every day, it is practically certain someone's pet *will* pick up a tick carrying Lyme disease. Similarly, if Alice visits the Gorge with Mollie three hundred times, Mollie is likewise certain to be exposed to a disease bearing tick.

Alice might feel differently about a beloved let getting Lyme disease than about the probability her car will start; in that case she may wish to adjust her level of practical certainty. This also brings in the complications of decision theory. If one thinks in terms of mundane lotteries (where neither positive nor negative outcomes have drastic consequences) rather than diseases, or the commonplace assumptions one makes going about daily business, the conclusions reflect common sense.

References

Allen, J., & Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial Intelligence, 15,* 143–178.

Brooks, R. (1990). Elephants don't play chess. *Robotics and Autonomous SYSTEMS, 6,* 3–15.

Calude, C. S. (2018). A probabilistic anytime algorithm for the halting problem. *Computability, 7,* 259–271.

Cangelosi, A., Bongard, J., Fischer, M., &, Nolfi, S. (2015). Embodied Intelligence. In *Springer Handbook of Computational Intelligence*, Kacprzyk, J., and Pedrycz, J., eds, (pp. 697–714).

Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown and Co.

Feldman, J., & Sproull, R. (1977). Decision theory and artificial intelligence IIL the hungry monkey. *Cognitive Science, 1,* 158–172.

Ford, K.M., &, Hayes, P. (1995). Turing Test Considered Harmful. In *Proceedings of IJCAI 1995,* pp. 972–977

Ford, K., Hayes, P., Glymour, C., & Allen, J. F. (2016). Cognitive orthoses: Toward human-centred AI. *AI Magazine, 36*(4), 5–8.

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

French, R. (2012). Dusting off the turing test. *Science, 336*(6078), 164–165.

Good, I. J., & Turing, A. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40*(3–4), 237–264.

Harnad, S. (2001). Minds, machines, and Turing: the Indistinguishability of Indistinguishables. *Journal of Logic, Language, and Information*, *9*(4), 425–445.

Horn, B. (1985). *Computer Vision*. Cambridge: MIT Press.

Kelly, S. (2017). *Endurance: A year in space, a lifetime of discovery*, Decle Edge.

Kleene, S. C. (1952). *Introduction to metamathematics*. Amsterdam: North-Holland.

Kyburg, H. E., Jr. (1974). *The logical foundations of statistical inference* (Vol. 65). Berlin: Springer Science & Business Media.

Kyburg, H. E., Jr. (1983). The reference class. *Philosophy of Science, 50*(3), 374–397.

Lanier, J. (2017). *Dawn of the new everything: Encounters with reality and virtual reality*. New York: Henry Holt and Company.

Levesque, H.J. (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, H.J. (2009). Is it Enough to get the Behavior Right? *Proceedings of IJCAI-09* 1439:1444.

Levesque, H. J. (2014). On our best behaviour. *Artificial Intelligence, 212,* 27–35.

Levesque, H. J. (2017). *Common sense, the turing test, and the quest for real AI: Reflections on natural and artificial intelligence*. Cambridge: MIT Press.

Leviathan, Y., & Matias, Y. (2018). Google duplex: An AI system for accomplishing real-world tasks over the phone. Google AI Blog. Retrieved March 11, 2019 from, https://ai.googleblog.com/2018/05/duplex-aisystem-for-natural-conversation.html.

Loui, R. P. (1987). Defeat among arguments: A system of defeasible Inference. *Computational Intelligence, 3,* 100–106.

Marcus, G. (2013). Why can't my computer understand me? *The New Yorker,* https://www.newyorker.com/tech/annals-of-technology/why-cant-my-computer-understand-me .

McCarthy, J. (1986). Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence, 28*(1), 89–116.

McCarthy, J. & Hayes, P.J. (1969) Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence* **4**, (eds Meltzer, B. and Michie, D.). Edinburgh: Edinburgh University Press 463–502.

McDermott, D., & Doyle, J. (1980). Non-monotonic logic. *Artificial Intelligence, 13*(1–2), 41–72.

Mori, M. (2012) The Uncanny Valley: The Original Essay by Masahiro Mori, translated by MacDorman, K.F., & Kageki, N. (2012). *IEEE Spectrum* ,pp. 1–8.

Neufeld, E., & Finnestad, S. (2016a). Artificial intelligence testing. In *Proceedings of the Twenty-Ninth International FLAIRS Conference*, pp. 158–161

Neufeld, E., & Finnestad, S. (2016b). The mismeasure of machines. In *Proceedings of the 29th Canadian Conference on Artificial Intelligence*, pp. 58–63

Neufeld, E., & Finnestad, S. (2016c). The Post-Modern Homunculus. In *Proceedings of the European Conference on Artificial Intelligence*, pp. 1670–1671

Neufeld, E., & Finnestad, S. (2020). In defense of the Turing test. *AI & Society*, to appear.
Neufeld, E., & Goodwin, S. (1998). The 6–49 Lottery Paradox. *Computational Intelligence, 14*(3), 273–286.
Núñez Siri, J., Neufeld, E., Parkin, I., & Sharpe, A. (2020). Using Simulated Annealing to Declutter Genome Visualizations. In *Proceedings of the Thirty-Third International FLAIRS Conference*, pp. 201–204
Poole, D. (1985). On the Comparison of Theories: Preferring the Most Specific Explanation, in *Proceedings of IJCAI 1985,* Morgan Kaufmann, pp. 144–147.
Poole, D. (1989). What the lottery paradox tells us about default reasoning (extended abstract). In *Proceedings of KR-89,* Toronto, pp. 333–340.
Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*, Allen Laine.
Russell, S. & Norvig, P. (2009) *Artificial Intelligence: A Modern Approach (3rd edition)* Pearson.
Searle, J. (1980). Minds, brains, and programs. *Brain and Behavioural Sciences, 3,* 417–457.
Shannon, C. & McCarthy, J. (1956). *Automata Studies*. Princeton University Press, p.vi
Shotter, J. (2019). Why being dialogical must come before being logical: the need for a hermeneutical–dialogical approach to robotic activities. *AI & SOCIETY, 34*(1), 29–35.
Trausan-Matu, S. (2019). Is it possible to grow an I-Thou relation with an artificial agent? A dialogistic perspective. *AI & SOCIETY, 34*(1), 9–17.
Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 2*(42), 230–265.
Turing, A. (1950). Computing machinery and intelligence, *Mind* 433–460.
Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence, 13*(1–2), 81–132.
The Guardian. (2014). *Computer simulating 13-year-old boy becomes first the pass Turing test*, https://www.theguardian.com/technology/2014/jun/08/super-computer-simulates-13-year-old-boy-passes-turing-test.
Van Arragon, P. (1991). Modeling default reasoning using default. *User Modelling and User Adapted Interaction, 1,* 259–288.
Wired Magazine. (2011). *Spide rSpins Zero-Gravity Web in Space*, https://www.wired.com/2011/06/space-spiders-action/.