Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Artificial Intelligence Testing

Eric Neufeld and Sonje Finnestad
Department of Computer Science, 110 Science Place
University of Saskatchewan, Saskatoon, Canada, S7N 5C9
eric.neufeld@usask.ca

Abstract

Hector Levesque has a strong critical position regarding the place of the Turing Test in Artificial Intelligence. A key argument concerns the test's use of, or even, reliance on *deception* for subjectively demonstrating intelligence, and counters with a test of his own based on Winograd Schemas that he suggests is more objective. We argue that the subjectivity of the test is a strength, and that evaluating the outcome of Levesque's objective test introduces other problems.

Introduction

A few years ago, Hector Levesque's critique of, and proposed alternative to, the Turing Test made the pages of the New Yorker – a fantastic achievement (Marcus, 2013). Levesque objects to the Turing Test because it involves *deception*: a computer must *deceive* or *fool* observers into thinking it is human. A related point is the 'trickery' and 'evasiveness' of the Loebner competition chatterbots: "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).

Levesque proposes instead an objective multiple-choice test based on Winograd Schemas, clever puzzles requiring anaphor disambiguation, which, he argues is less subject to abuse. Contests based on Winograd Schemas are already being planned (Morgenstern and Ortiz, 2015).

Levesque makes a good case, but we respectfully disagree. First, a machine that could *consistently and over time* pass as a human engaging in ordinary conversation would be a mechanical wonder. Secondly, though meeting Levesque's challenge would be a remarkable technical feat, we have questions about the scoring algorithm; neither are we convinced this would demonstrate intelligence superior to, say, IBM's Watson.

The Turing Test

Turing (1950) first describes the Imitation Game as a game played by a man, a woman, and an interrogator of either sex. The interrogator, who is in another room, must determine which of the other two players is the man and which is the woman.

"We now ask … ," Turing says, "'What will happen when a machine takes the part of [the man] …?' *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?*" (our italics) The italicized question replaces the original, 'Can machines think?'

A year or so later, in 1951, in a lecture broadcast on BBC radio as, "Can Digital Computers Think?", Turing describes another version of the test as "something like a viva-voce examination."

Yet another variant of the test appears in a script of a discussion, "Can Automatic Calculating Machines Be Said to Think?", presented on BBC radio in 1952. This version of the test involves a 'jury' of interrogators and successive trials in which the jury questions a hidden man or machine (they know not which) and must render judgment as to whether they are talking to man or machine.

In the same discussion, Turing describes 'the idea of the test': "The idea of the test is that the machine has to try and pretend to be a man, by answering questions put to it, and it will only pass if the pretense is reasonably convincing." He added that it would be at least 100 years before the machine would "stand any chance with no questions barred" (Turing, 1952, p 495).

The Gender Game

What is the significance of the gender game? Why does Turing tell this odd little story before getting to the test in its familiar form?

Though Hodges, calls it a 'red herring' (Hodges, 2012, p. 415), Copeland sees it as key, "part of the protocol for

scoring the test": can a computer imitate a human being as well as a man can imitate a woman (Copeland, 2004, p. 436)? The gender test, we argue, is not a red herring and, if it is a scoring protocol, it is more than that: the gender game is an interpretive guide.

We claim that, socially, there are different cultural norms for men and women - perhaps less so now than in 1950, but still the case. Such norms are comprehensive, complex, often subtle, and variable. For a man to pass as a woman, even via teletype, requires not only a deep and extensive familiarity with the norms, but also the ability to 'inhabit' them in ordinary conversation, and thereby 'deceive' a judge – but this 'deception' is different from the Loebner chatterbot behaviour Levesque objects to.

But is it likely, one may counter, that in the England of the 1950's, well before the advent of second wave feminism, this is the kind of thing Turing had in mind? We do not think this implausible, for two reasons.

First, Turing was, in the terminology of the time, homosexual. He was, particularly where sexuality and gender were concerned, an outsider and he would have experienced many social norms, and – significantly – norms associated with gender, *from outside* (MacCulloch 2013). He must have been acutely aware of the existence and power of these norms; he would have bumped up against them throughout his life, not just in the last years, when his nonconformity had such damaging consequences.

Secondly, there is good reason to believe that Turing, had he been aware of it, would have placed himself somewhere on the autistic spectrum (O'Connell and Fitzgerald, 2003). Many 'high-functioning' autistic people describe an experience of social norms from the outside. Such norms, they say, are not part of (or internal to) them; they are neither instinctive nor taken for granted but learned, often slowly and painfully, through a prolonged process of careful social analysis and no little negative reinforcement. Temple Grandin describes herself as "an anthropologist on Mars" (Sacks, 1995).

The Gender Game and the Turing Test

Turing introduces his test by inviting his overwhelmingly male readership to imagine what it would be like to imitate a woman under the conditions given, that is, in conversation via teletype. Now, he says, consider a game wherein a computer must imitate a human being.

If the gender test is a test of how well, or how convincingly, a man can follow the norms of what might be called 'female culture', considered more generally, it is a test of how well someone can follow the rules – or mores – of another culture, rules that inform the lives of *others*, rules that are *outside of* or *external to*, as opposed to *instinctive* or *internal to*, him or herself, and must be learned. A good way to understand what is involved in thinking and speaking – or carrying on a conversation – is to experience or at least (as in this case) to imagine what is involved in thinking and speaking in terms of a culture not our own. This is the kind of thing, Turing says, the computer is required to do.

The imitation game, so understood, is a test of how well a computer can follow norms of intelligent behaviour, as expressed in conversation (via text). Such rules, of course, are always historically and culturally specific, which means that a Turing test will always be culturally specific; there can be no universal Turing test (e.g. Cohen, 2005). We maintain that this is no argument against the test; indeed, on this understanding of the gender example, it is, at least implicitly, part of Turing's point. (We also note that the Winograd Scheme Challenge Levesque advocates is likewise culturally specific.)

It is from this perspective that we shall address Levesque's critique of the Turing test and his proposal for a new kind of test.

The Turing Test and Deception

Deception is part and parcel of Turing's test. If a computer is to imitate or simulate a human, it is going to have to lie or, at least, communicate untruths! There's no getting around that. But is it a problem?

According to Levesque, it's a problem big enough to warrant ditching the test. "The Turing test," he says, "has a serious problem: it relies too much on *deception*. A computer program passes the test iff it can fool an interrogator into thinking she is dealing with a person not a computer." This means that a program "will either have to be evasive … or manufacture some sort of false identity (and be prepared to lie convincingly)" (Levesque, 2014). Elsewhere (Levesque, 2011), he refers to "some troubling aspects" of the test and mentions first, "the central role of deception".

This is not the way Turing describes his test; this is loaded language and there is, perhaps, a whiff of moral disapproval in all this. It is hard to know what to do with a moral objection in a case like this, if indeed that is the nature of the objection: it seems out of place. We do not claim that ethics has nothing to do with AI or with science in general but neither are we discussing fraud or killer drones; we are talking about a 'test' and, indeed, a test that its creator frequently describes as a 'game'.

(We note that Levesque does say, at one point, that, "it's not really lying" since it's "just a game" (Levesque, 2011); nonetheless, a significant portion of his critique is devoted to the place of deception in the Turing Test.)

Human Interaction and Deception

It is not just the nature or, we might say, the norms, of the game *qua* game that call this critique into question but the norms of this *particular* game, which is a *cultural/social-norm-following game*. This is so because any participation in a culture or sub-culture is going to involve a certain amount of what might be described as deception: e.g., inquiries into the well-being of others, adherence to dress codes, employee compliance with the oft-legislated duty of loyalty and appropriate workplace behaviours, and so on.

More to the point, perhaps, since what the computer is required to do might be considered a species of acting or performative fiction, a complex set of norms governs the production and consumption of cultural artifacts like fiction, poetry, theatre, film, and tv, all of which involve interplay between truth and imaginative construction.

In sum, various forms of what might be described as deception are pervasive in human societies and failing to practice or appreciate them is apt to make people suspect one's intelligence (though of course there may be other reasons). Again, then, we wonder why the deception involved in the Turing test – which is a test for intelligent behaviour - should be considered such a problem.

The Issue

The real issue, surely, is whether the deception involved in the test in the ideal undermines it as a test of intelligence. Levesque sometimes seems to suggest that this is so, although this suggestion seems to belong to another objection, this one to the conversational form of the test: "The conversational format of the Turing Test allows a cagey subject to hide behind a smokescreen of playfulness, verbal tricks, and canned responses" (Levesque, 2014). Consider, he says, Eliza. Consider the Loebner competition chatterbots, he says.

Free form conversations, Levesque acknowledges, are "the best way to get to know someone, to find out what they think about something, and therefore *that* they are thinking about something" (Levesque, 2011). We agree that attempted distractions and evasions, while they may in some contexts appear to demonstrate some intelligence, distract and detract from the goal or aim of Turing's test.

Yet in practice such behaviour actually demonstrates a *lack* of intelligence. To wit, neither Eliza nor the chatterbots can sustain the impression of intelligence.

Surely this is not something Turing intended. The Loebner Competition is a contest; the primary goal, for the contestants, is to win; the goal of the Turing Test, on the other hand, is to demonstrate and assess intelligent behaviour. The differences are such that we question whether the behaviour of the Loebner chatterbots, however outrageous, even with Eliza for extra weight, can be a convincing argument against the Turing Test.

Levesque seems to say, further, that a machine should be able to show us that it is thinking without having to "fool an interrogator into thinking she is dealing with a person not a computer" (Levesque, 2014).

We note first that Turing did not say that his was the only possible, valid, or useful test. It is not that a computer *must* pass as a person to demonstrate its intelligence; it is that if a computer *can* pass as a person, we must consider it intelligent.

It is a great strength of Turing's test that it does not depend on a definition of intelligence or thinking, whether human thinking, machine thinking, or thinking in general. Turing asks, in effect, "how do we recognize intelligence in other people?" and affirms that "if we see this in a machine, we will have to say that it, too, is thinking". Even if there are other tests and even if machine thinking is very different from human thinking in any form, if a machine can pass the test, that is, if it can successfully emulate intelligent human behaviour, we have as much reason to ascribe intelligence to it as to anyone or anything else. The Turing Test, it might be said, is our everyday test for intelligence applied to machines: "we know it when we see it". Whatever else might or might not be possible or, for one reason or another, desirable, this is surely of value.

Levesque's 'New Type of Turing Test'

Levesque suggests that "what we are after is a new type of Turing test" and proposes, as an alternative, a *Winograd Schema Test*. A Winograd Schema (WS) "is a small reading comprehension test involving a single binary question" (Levesque, 2011). For example,

> The trophy would not fit in the brown suitcase because it was too big. What was too big?
>
> Answer 0: the trophy
> Answer 1: the suitcase.

A WS must have certain features, one of which is the presence of a *special* word that, when replaced by an *alternate* word, flips the answer. (The special word in the above example is *big* and the alternate word is *small*.)

These are undeniably ingenious questions, but a WS test is not without its own problems. There is, for example, the matter of scoring. What is a passing grade? Levesque does not say. (Interestingly, he makes a similar charge against the Loebner competition: "Grading the test … is problematic" (Levesque, 2011).)

This is complicated by the fact that a machine can do some guessing, including some educated guessing, and get a great many right answers yet make a telling wrong an-

swer, as happened when Watson gave "Toronto" as the answer to a question from the category "US cities".

We have been challenged on this argument, on the grounds that any reasonable scoring algorithm *can* be a measure of intelligence. We are of the view that, although Watson scored significantly higher than all of the human players, for most observers, the illusion of *intelligence* vanished with this single wrong answer. To say that the scoring algorithm is not important, in our view, regresses the philosophical issue of true machine intelligence to the scoring algorithm – which replaces Turing's human judge(s).

With a scoring algorithm, will the Levesque test be equivalent to the Turing test? Throughout the history of AI, scholars who regarded the Turing test as the ultimate test, but beyond the reach of the machines of their era, threw out other milestone problems as testbeds: chess and checkers, where the domain assumptions were very simple yet the search space inconceivably large. Checkers (Schaeffer *et al*, 2007) has turned out to be a solved game, but one that still requires considerable resources to play perfectly. It does not seem far-fetched to expect chess to be solved, even if it is in the distant future. At one point, and for good reason, scholars believed that such games could only be played well by machines if those machines were intelligent in some sense. With the passage of time, it appears that these games are being played well by machines with highly developed (or nearly complete) 'answer books' – much harder than, but along the same principles as, tic-tac-toe. But does this constitute intelligent behaviour? We think most scholars would say no.No exceptions.

Conclusions

There is, we maintain, no substitute for the Turing test, at least so far. Turing compared his test to judgment by a jury. The analogy is apt in more ways than one.

There is no agreed-upon definition of justice any more than there is of intelligence. When it comes to the determination of justice in particular cases, we rely on judges and juries to decide. We can't precisely define how this is done any more than we can precisely define justice.

Interestingly, there are at least two theories of why juries exist (Burns, 1995). The one most extant is that juries of peers tempered the decisions of judges, in the same sense the introduction of the House of Commons tempered decisions of the House of Lords. The other might be characterized as saying that the idea of justice ultimately resides in the minds of humans.

The jury system, like the Turing Test, has been subject to a variety of criticisms, many of them apparently rather devastating. Yet in spite of our awareness of imperfections and abuses, most observers would not want to replace something of such value unless and until its fatal defects and the superiority of the proposed alternative are convincingly demonstrated.

Thus far, this has not happened; neither, we maintain, has it happened in the case of the Turing test. At this point, to quote Levesque in another context (Levesque, 2011), the Turing test is "the best game in town", and its subjectivity is its strength.

References

Burns, R.P. 1995. The History and Theory of the American Jury. *California Law Review* 83(6):1477-1494.

Cohen, Paul R. 2005. If not Turing's test, then what? *AI Magazine* 26(4):61-67.

Copeland, B.J. 2000. The Turing Test. *Minds and Machines* 10(4):519-539.

Copeland, B.J., ed. 2004. *The Essential Turing*. Oxford University Press.

Hodges, Andrew. 2012. *Alan Turing: The Enigma*. Random House.

Levesque, H.J. 2009. Is it Enough to get the Behaviour Right?" In *Proceedings of IJCAI-09, Pasadena, CA*, Morgan Kaufmann, 1439:1444.

Levesque, H. J. 2011. The Winograd Schema Challenge. In *Logical Formalizations of Commonsense Reasoning, 2011 AAAI Spring Symposium, TR SS-11-06*.

Levesque, H.J. 2014. On our best behaviour. *Artificial Intelligence* 212(1): 27-35.

MacCulloch, Diarmaid. *Silence: A Christian History*. Penguin, 2013.

Marcus, G. 2013. Why Can't My Computer Understand Me? *New Yorker*, August 14, 2013.

Morgenstern, L., and Ortiz, C. 2015. The Winograd Schema Challenge: Evaluating Progress in Commonsense Reasoning. In *Proceedings of Innovative Applications of Artificial Intelligence 27th IAAI Conference,* 4024-4025.

O'Connell, H., and Fitzgerald, M. 2003. Did Alan Turing have Asperger's syndrome? *Irish Journal of Psychological Medicine* 20(01): 28-31.

Sacks, O. *An Anthropologist on Mars: Seven Paradoxical Tales*. Alfred A. Knopf. 1995.

Schaeffer, J., Burch, N., Björnsson, Y., Kishimoto, A., Müller, M, Lake, R., Lu, P., and Sutphen, S. 2007. Checkers is solved. *Science* 317(5844):1518-1522.

Turing, Alan M. "Computing machinery and intelligence." *Mind* (1950): 433-460.

Turing, Alan. 1951. Can Digital Computers Think? In *The Essential Turing,* B. J. Copeland, Ed., Clarendon Press, Oxford, 2004: 476-486.

Turing, Alan, Richard Braithwaite, Geoffrey Jefferson, and Max Newman. 1952. Can Automatic Calculating Machines be said to Think? In *The Essential Turing,* B.J. Copeland, Ed., Clarendon Press, Oxford 2004: 487-506