Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest

Luciano Floridi · Mariarosaria Taddeo · Matteo Turilli

Abstract An evaluation of the 2008 Loebner contest

Keywords AI · Loebner contest · Turing test

L. Floridi (✉)
Department of Philosophy, University of Hertfordshire, Hatfield, UK
e-mail: l.floridi@herts.ac.uk

L. Floridi
Faculty of Philosophy, University of Oxford, Oxford, UK

L. Floridi · M. Taddeo · M. Turilli
Information Ethics Group (IEG), University of Oxford, Oxford, UK

L. Floridi · M. Taddeo · M. Turilli
Research Group in Philosophy of Information (GPI), University of Hertfordshire, Hatfield, UK

M. Taddeo
International Society of Ethics and Information Technology (INSEIT),
University of Wisconsin-Milwaukee, Milwaukee, USA

M. Taddeo
Department of Philosophy, University of Hertfordshire, Hatfield, UK

M. Taddeo
Department of Philosophy of University of Padua, Padua, Italy

M. Turilli
OUCL, University of Oxford, Oxford, UK

M. Turilli
Oxford e-Research Centre (OeRC), University of Oxford, Oxford, UK

M. Turilli
Centre for Ethics and Economics and Business, Universidade Católica Portugueasa,
Lisbon, Portugal

This year, the Loebner Prize competition[1] was held in England, at the University of Reading to be precise.[2] It was paralleled by a Symposium on the Turing Test (henceforth TT, Turing (1950)) organised by the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB). Expectations were high, and very highly advertised too. Kevin Warwick, the organiser (together with Huma Shah), seemed to believe that this might well be the time when machines would pass the TT: "The competition is all about whether a machine can now pass the Turing Test, a significant milestone in Artificial Intelligence. I believe machines are getting extremely close—it would be tremendously exciting if such a world first occurred in the UK, in Reading University in 2008. This is a real possibility".[3] Having been invited to play the role of judge, together with several colleagues, we were indeed very excited but also very sceptical. We doubted that machines could pass even a simplified TT. In any case, we enjoyed the opportunity to see close-up the organisation and machinery of the TT. It was intriguing.

The test was organised as follows: four judges were each asked to sit in front of a computer in order to interact with two interlocutors at the same time by means of a chat system, very similar to the—by now well known—IM clients. Following Turing's suggestion, the interaction was limited to 5 min altogether (this meant an average of 2.5 min dedicated to each interlocutor on the split screen), at the end of which every judge was asked to declare whether one, both or neither of the interlocutors were machines. This feature was a confusing departure from the original Turing Test, which we discovered only because one of us (Matteo Turilli) became convinced that both interlocutors were computers (he was right). The whole process was then repeated by shifting every judge to a new station in order to make them test a total of four pairs of interlocutors. Each time, the judge was asked to score the machine capabilities between 0 and 100, or to guess the gender, age, speaking abilities (e.g. native speaker) and potential linguistic impairments of the human interlocutor.

As we had expected, and despite the brevity of our chats, a couple of questions and answers were usually sufficient to confirm that the best machines are still not even close to resembling anything that might be open-mindedly called vaguely intelligent. Here are some convincing examples.

One of us (Luciano Floridi) started his chat by asking: "if we shake hands, whose hand am I holding?" One interlocutor, the human, immediately answered, meta-linguistically, that the conversation should not have mentioned bodily interactions. Indeed, he later turned out to be Andrew Hodges, Turing biographer (Hodges 1983), who had been recruited on the spot in order to interact with one of us (LF) on the other side of the screen. On the other hand, the computer, which turned out to be "jabberwacky",[4] failed to address the question and spoke about something else, a trick used by many of the tested machines: "We live in eternity. So, yeah, no. We

[1] http://www.loebner.net/Prizef/loebner-prize.html retrieved 12 November 2008.

[2] http://www.reading.ac.uk/cirg/loebner/cirg-loebner-main.asp retrieved 12 November 2008.

[3] http://www.reading.ac.uk/sse/about/News/sse-newsarticle-2008-05-16.asp retrieved 12 November 2008.

[4] http://en.wikipedia.org/wiki/Jabberwacky retrieved 12 November 2008.

don't believe." It was the usual, give-away, tiring, Eliza-ish strategy (Weizenbaum 1966), which we have now seen implemented for decades. Yet another confirmation, if one was still needed, that while a dysfunctional pseudo-semantic behaviour like Eliza's or Parry's could fool some human interlocutor in a highly specific context, it is utterly unsuccessful in a general purpose, open conversation.

The second question did not change the situation, but merely confirmed the first impression: "I have a jewellery box in my hand, how many CDs can I store in it?" Again, the human interlocutor provided some explanation, but the computer blew it badly. More Eliza. The third question came at the end of the 5 min: "The four capitals of the UK are three, Manchester and Liverpool. What's wrong with this sentence?" Once again, the computer went bananas.

All the other conversations developed rather similarly, although we (and especially Mariarosaria Taddeo) posed a different range of questions, using elementary logic ("if London is south of Oxford, is Oxford north of London?"), common shortcuts ("do *u* like to have dinner *b4 u* go to the cinema?"), figures of speech ("how does the colour red smell?") and even simple, plain enumerations ("could you tell me three things that you could do with a telephone?"). All these questions immediately gave away both humans and machines, making it unnecessary for any further interaction or tests, such as connecting multiple questions, "remembering" previous answers, or revising previous statements on the basis of new evidence.

If the TT at Reading went less badly than it could have (some machines did manage to fool some judges a few times), this is probably because some of the judges were asking useless questions, like "are you a computer?" or "do you believe in God?" (these are real instances). This was a sign that two essential points of the whole exercise had been missed by them (the judges, not the machines).

First, and especially given the very short interaction, answers should be as informative as possible, which means that one should be able to maximise the amount of useful evidence obtainable from the received message. It is the same rule applied in the 20 questions game: each question must prompt an answer that can make a *very significant* difference to your state of information, and the bigger the difference the better. But in the examples above, either "yes" or "no" will leave you absolutely unenlightened as to who your interlocutor is, so that is a wasted bullet.

Second, questions must challenge the syntactic engine which is on the other side. So other questions such as "what have you been up to today?" or "what do you do for a living?" (again, two real examples) are rather useless too. The more a question can be answered only if the interlocutor truly understands its meaning, context or implications, the more that question has a chance of being a silver bullet.

Two documentaries by the BBC[5] show both points been badly overseen by the judges. But if all the judges had followed this simple "*vademecum* for a TT judge",

[5] http://news.bbc.co.uk/2/hi/technology/7666836.stm and http://www.bbc.co.uk/berkshire/content/articles/2008/10/12/turing_test_feature.shtml both retrieved 12 November 2008.

we suspect that their first question would have almost always been sufficient to discriminate between the human and the machine. It certainly was for us.

Seven speakers took part in the parallel Symposium organised by Mark Bishop: Margaret Boden, Selmer Bringsjord, Susan Greenfield, Andrew Hodges, Owen Holland, Michael Wheeler and one of us (LF). Several participants defended the view that a serious TT would have to last much longer than a handful of seconds. We agreed, but we would also contend that this is as much because of the examined agents, and the slow means of communication (you have to write/read everything on a screen), as because of the judges, and their lack of training. If you need to test, and we mean really *test*, an artefact, the higher the stakes are, the tougher the procedure should be. We do not adopt the same standards when it comes to testing the safety of a house's central heating system or the safety of a nuclear power station. Why (artificial) intelligent behaviour should be tested by the untrained, naïve and often uninformed "man in the street" remains a mystery to us, pace Turing's suggestion. Unless that is the sort of dude you wish to fool.

It might be that the Loebner Prize should be re-thought more like a chess tournament, where we could play imitation games with different levels of time control: long games (up to 7 h), short games (30/60 min), blitz games (3–15 min for each player), bullet games (under 3 min) and one-question games (1 min). Almost all the computers we tested could not even pass the latter. Our score for them was zero.

The AISB Symposium was meant to provide plenty of food for our natural minds. We enjoyed the lively interactions, and found the presentations interesting and informative. To give an example, Margaret Boden suggested in her talk that we look at the TT from a different perspective. Her idea is to use it to assess the artistic skills of machines instead of their natural language competences. The suggestion is intriguing, and so were the examples of painting and music writing machines that she described. While we have no problem in seeing how a painting or musical piece (re)produced by those machines could easily be compared to a human artistic artefact (we would recommend to the sceptic the extreme case represented by Fontana's famous series of cut-off paintings entitled "quanta"), we are more sceptical about the idea that such machines are indeed imitating human artistic skills at all. We see a fundamental difference between them in terms of autonomy and hence choice capabilities. A machine, or better its driving software, does not make choices, it simply executes a set of instructions or, in the case of random operations, it effectively tosses a coin to make its selection. When we avoid catchy metaphors, computer programs that paint or output a score appear for what they are: very complex high-tech versions of the old paint brush or goose quill. They are instruments in the hand of the artist, not artists made by engineers. The problem is the usual one and it affects not only supposedly 'artistic' machines but all software programs that seem to exhibit some sort of intentional skill: their semantic capabilities are in our eyes not in their codes.

We found the first half of Owen Holland's talk about the Ratio Club particularly inspiring. In his presentation, he recounted the story of this small, informal dining club of young researchers (professors were explicitly banned) founded in 1949 (its last meeting was in 1958). Some of the leading psychologists, physiologists, mathematicians and engineers met over dinner to discuss issues in cybernetics.

William Ross Ashby and Turing himself were among its members (Husbands and Holland 2008).

We disagreed with several people, however, about the following issue, discussed during the second half of Holland's talk. There seemed to be some coalescing consensus on the view that a machine will pass the TT only if it is conscious. This is certainly not the case. The TT is a matter of semantics and understanding. Although we might never be able to build truly intelligent machines—as we suspect—consciousness need not play any role. This is not to say that a conscious machine would not pass the TT. For it would, of course, insofar as consciousness presupposes intelligence. Nor is it to say that smart applications will never be able to deal successfully with semantic problems by intelligence-free means. Some already do. Isn't it handy that Google knows better and tells you that your keywords are misspelled and should be so and so? But then, our dishwashers need no intelligence (let alone consciousness) to do a better job than we. What it does mean is that, after half a century of failures and more or less zero progress, some serious reconsideration of the actual feasibility of true AI is a must, and making things immensely more difficult can hardly help (although it might give some breathing space to a dying paradigm). To our philosophical astonishment, however, the argument seemed to be that, since we do not have the faintest idea about how to build a machine that can answer a handful of intelligent questions or even win the one-question TT consistently, the best strategy might be to go full-blown and try to build a machine that is conscious. As if things were not already impossibly difficult as they stand. This is like being told that if you cannot make something crawl, you should make it run a hundred metres under 10 s, because then it will be able to crawl. Surely it will, but surely there must be better ways of spending our research funds than by chasing increasingly unrealistic dreams. The fact that nobody agrees on what consciousness is can only help by mudding the water and making cheating easier. After all, if anything may count as consciousness, the game becomes somewhat easier. Turing, of course, knew better. He refused to define intelligence, so we should follow his advice and perhaps adopt a test for consciousness. One of us (Floridi 2005) has provided one in this journal, but others can be devised.

The day ended with the announcement that no machine had passed the TT. As usual, there was a winner of the Loebner's consolation prize for being the least disappointing machine. This was the programme Elbot,[6] created by Fred Roberts, who was awarded the $3,000 Loebner Bronze Award. He managed to convince three of the 12 interrogators that he was human. We agreed that it deserved the Prize more than the others.

References

Floridi, L. (2005). Consciousness, agents and the knowledge game. *Minds and machines, 15*(3–4), 415–444. doi:10.1007/s11023-005-9005-z.

Hodges, A. (1983). *Alan Turing: The enigma*. London: Burnett Books.

[6] http://en.wikipedia.org/wiki/Elbot retrieved 12 November 2008.

Husbands, P., & Holland, O. (2008). The ratio club: A hub of British cybernetics. In P. Husbands, M. Wheeler & O. Holland (Eds.), *The mechanical mind in history*. Cambridge, MA: MIT Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460. doi:10.1093/mind/LIX.236.433.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. doi:10.1145/365153.365168.