Kalissa Feinberg
December 7, 2019
Great Ideas in Computer Science

Almost 70 Years Later: Important Moments in the Turing Test

Kalissa Feinberg

Introduction[5]

In 1950, Alan Turing posed a bold question: "Can machines think?" Although it seems undefinable, Turing proposed a concrete technique: The Imitation Game (IG). In this contest, a chatbot and a man each masquerade as a man to try to convince a judge that they are the real man. Turing argues that if a judge cannot identify the real man with better than 50% accuracy, then we must say that the computer can "think."

Although we naturally distinguish our speech and our cognitive ability to "think," Turing fights against this notion. He argues that we assume that other humans around us "think," but we have no way of proving this other than the conversations we have with them. Thus, to prove that machines "think," we should require no evidence besides their ability to hold reasonable conversations with us.

In this paper, I look at what I perceived to be four of the most important moments surrounding the Turing Test since its inception. I look at both the development of chatbots and public perception of the test: it is impossible to separate the two from one another, since chatbots directly impact public perception and public perception directly impacts the types of chatbots created.

Turing's Proposal[5]

In the 70 years since Turing published his paper, one looming question persists: how do we program a computer to pass the Turing test?

Turing himself first attempted to answer this question. In line with his focus on human intelligence, Turing believed that a viable approach would be to model human learning abilities. He suggested we simulate a child's mind – a blank slate with an ability to learn – then train it with massive quantities of data. He also suggested a more physical approach: replicate human senses in a computer and teach it to speak and understand English the way that a child would be taught.

Although this aspect of his paper proposes an "honest" solution to the task at hand, Turing was not opposed to the use of deception in the IG, the same way that a man might use deception in pretending to be a woman. One example that Turing provides is that a judge asks the computer: "Add 34957 to 70764." Although the computer could respond to this question instantly, it pauses 30 seconds before responding: "105621." Turing does not provide one clear path he believes will be useful to winning his game; he seems to believe that producing human thought will require a combination of genuine learning, like a child's, and tricks, like slowing down computations to make them more "humanlike."

ELIZA[6]

One of the first, and most influential, chatbots to take a stab at Turing's Test was ELIZA. Developed by Joseph Weizenbaum in 1966, ELIZA was framed as a kind of psychotherapist bot. For example, a conversation with ELIZA could go something like this:

Participant: Men are all alike.
ELIZA: In what way?
Participant: They are always bothering me about something or other.
ELIZA: Can you think of a specific example?[6]

And so on. Researchers taught ELIZA to communicate through rules and keyword associations. For example, ELIZA cycles through inputted blocks of text like "what makes you think…" and automatically translates inputted words – "you" translates to "me," "my" translates to "your," etc. When forming a response, ELIZA parses through the inputted sentence for keywords to use in the response, rates each keyword by importance and generates a response using the most important keyword that it identified. In the case that it is unable to respond to a given sentence, ELIZA retains a fraction of the prior conversation in its memory so it can respond with something like "tell me more about X you were saying earlier."

Despite its simple design, ELIZA performed surprisingly well at the Turing Test. Many people believed she was a human, and some even formed "emotional attachments"[4] to her. It differed, however, from Turing's initial vision of a "thinking" AI in a number of ways.

For one, ELIZA did not implement any of the "general" childlike learning that Turing foresaw, but followed basic rules instead. It could perform in a limited Turing Test as a therapist, but could not take on any other personas.

In explaining how ELIZA processes information, Weizenbaum actually compares the processes to a "foreigner" that knows only a few words of English. To perform understanding, the "foreigner" can pick out the words that they understand and manipulate them into coherent English without understanding the topics they were discussing.

Loebner Competition[1]

In 1950, Turing believed that in 50 years, machines would be able to pass the Turing Test.[5] In 1990, researchers were not close to achieving this goal, but Hugh Loebner proposed a contest: an annual competition in which chatbots would compete to pass the Turing Test. The first winner would earn $1500. If a chatbot actually passed the Turing Test, the developer would win $100,000 and the competition would never be held again.

The competition used the framework of Turing's original test, but with some slight variations. For one, the organizers replaced the two-terminal design. Instead, the competition's ten judges were told they would chat with at least two humans and at least two bots. Rather than simple binary guesses between computer and man, the judges instead rated each terminal on a 100-point scale from least human to most human. To maintain some binary, they marked a dividing line in their rankings between men and chatbots. This allowed the competition to maintain Turing's original binary, while also establishing a clear ranking of the bots to determine a winner of the competition.

In some ways, the competition continued Turing's original vision. Although bots didn't directly complete with humans, there was a clear human-bot binary. As Turing suggested, bots were allowed to perform "tricks," such as making fake typos or typing at a "human" pace.

Unlike Turing's original vision, however, the competition organizers decided that the initial competition would be limited – judges were restricted discussing only certain topics with each terminal, in the same way that conversations with ELIZA were limited to a "therapist"-type environment. Additionally, conversations were time restricted – judges had around only five minutes to determine if the bot was human or not. Although Turing never specified any clear time window in his test, it seemed as though the judge would continue asking questions until they believed they

had distinguished the human and bot. The winner in 1991, "Whimsical Conversation" was a bot that performed like a jester:

> **01:33:03 Terminal 5** Hello there judge, are you ready to have some fun? How ya doin, judge? Hello—how's the weather up there today?
> **01:33:35 Judge 6** Dark clouds with occasional bursts of sunshine (or rather camera flashes).
> **01:34:45 Terminal 5** The trouble with the rat race is that even if you win you're still a rat. Whimsy can be flimsy, right? How important is that to you?

"Whimsical Conversation" was sometimes nonsensical in its conversations, like in its "rat race" statement, but judges overlooked this as fitting the "jester" persona and the terminal conversation topic of "whimsical."

The competition organizer drew a comparison between "Whimsical Conversation" and ELIZA. Both apply simple tricks: planned stories, like Whimsical Conversation's "rat race" or ELIZA's set phrases at the beginning of its sentence. Both interact in a stilted way: ELIZA asks only questions or repetitions of the initial phrase while Whimsical Conversation will ignore the previous phrase and brings up a completely different topic. They do so in a kind of opposite manner, however: whereas ELIZA cannot converse about topics outside of what the human chooses, "Whimsical Conversation" struggles to clearly converse about the judge's topic of choice. "Whimsical Conversation," however, improves on ELIZA's error recovery:

> **Judge** Are you repeating me just for laughs?
> **01:38:04 Terminal 5** I'm sorry, I do sometimes repeat myself...perhaps it's because i can't find my electr]nic organizer!.

Not only does "Whimsical Conversation" apologize when it perceives the judge to be angry, it also uses faux-typos to further maintain a pretense of humanity.

The first major takeaway from the Loebner competition was that we were further along in our progress with chatbots than we had originally anticipated. The second was that bots like ELIZA and "Whimsical Conversation," which behave as a specific persona in a specific scenario, had more success so far than any kind of "all-purpose" bot that Turing may have originally envisioned.[4]

Pushback to the Competition[2]

Despite its media hype and financial incentive, the Loebner competition was derided by experts in the field. In 1995, two professors, Patrick Hayes and Kenneth Ford, wrote a letter explaining why they believed the Loebner competition and the Turing Test were not only useless to the field of AI, but also could be actively harmful.

For one, they took issue with the Loebner competition's acceptance of "tricks," like "Whimsical Conversation's" misspelling of words, or bots purposefully typing and deleting at a humanlike rate. They argued that these computer programs did not require any breakthroughs or important discoveries, just the ability to encode simple manipulations that deceived judges.

Turing and the organizers of the Loebner competition had a grand vision for what chatbots could do once they reached Turing-test level competency. Robert Epstein, one of the organizers, envisioned a world in which efficient, natural language interfaces would be able to parse massive amounts of data and communicate it to us in simple human language. He even goes so far as to suggest that "thinking computers will be a new race, a sentient companion to our own,"[1] and suggests that one day, when the human race has died out, the mechanical race will view us as deities because we are their "creators."[1]

In reality, Hayes and Ford argue, not only is this kind of goal unrealistic, but also it is not the kind of bot that the Loebner competition rewards. Bots like ELIZA that can perform well in the

Turing test are not necessarily useful to the field of AI. In fact, if we did produce humanlike intelligence, what would be the point? We already have more than enough humans in the world.

Instead, Hayes and Ford view AI as a tool to assist humans. Rather than create a kind of all-purpose AI with general human intelligence, they believed we should continue to focus on the kind of "idiot savant"[2] AI that excels in one specific area.

Turing's initial paper was anthropomorphic, which makes sense given its context: we were the only kind of intelligence we were familiar with, so of course this we would try to emulate ourselves. ELIZA and "Whimisical Conversation" both attempted to emulate one specific type of person, and this was why they had success in the Turing Test.

Epstein viewed the Turing Test as an opportunity to one day develop a "companion" to humans.[1] Hayes and Ford, however, argue that this type of view is outdated. They argue that when planes were first developed, we tried to model them after birds. We included flapping wings and even beaks because we knew birds could fly but didn't understand how, so we tried to copy anything we saw about them. In his initial paper, Alan Turing did the same with intelligence – attempt to model it after human intelligence. This, however, led to bots like ELIZA. Rather than attempt to model AI after humans completely, Hayes and Ford argue that we should instead simply use humans as inspiration for building something new entirely.

A New Test[3]

In recent years, rather than scrap the Turing Test altogether, numerous researchers have attempted to create a new, more useful type of Turing Test. In 2011, three researchers suggested one such test, the Winograd Schema challenge, as a viable alternative.

Like Hayes and Ford, the researchers argued that the Turing Test required too great a level of deception to be useful: what is the point of an AI determining a fake identity with random information like their height, parents, friends, etc? They pushed back against Loebner competition winners, like "Whimsical Conversation," that used fake or canned dialogue to impress the judges in a short time window.

Instead, they suggested a simple test with a few stipulations. Given a sentence and a question about the sentence, the computer must correctly answer the question. One such example would be the question below:

Joan made sure to thank Susan for all the help she had given. Who had given the help?

Answer 0: **Joan**

Answer 1: **Susan**

A human could answer this question immediately – clearly Susan has given the help, which is why she is being thanked by Joan. However, this is sentence is surprisingly difficult for a computer: both names are female, and it technically is possible that Joan could thank Susan for Joan's own help – we just use common sense to rule out this possibility.

This question is also not Google-able: it requires both the natural language processing ability to parse out the meaning of this sentence and the background knowledge of the world to identify the answer. Although this test lacks the elegant simplicity of the Turing Test, it provides solutions to some of the problems that researchers have identified with the Turing Test over the years.

Conclusion

This paper traced the development of a new form of Turing Test, the Winograd Schema challenge, back through criticisms of the Turing test, the creation of the Loebner Competition, and the development of one of the earliest well-performing chatbots. Through this timeline, we see that

as chatbots performed better on the Turing Test, criticisms of the Turing Test were raised, often based on the quality of chatbot that was able to pass. Although the Turing Test has been valuable in developing the chatbots we have today, as we learn more about natural language processing and AI, perhaps the Turing Test show grow with us.

Citations

1. Epstein, R. (1992), 'The Quest for the Thinking Computer', AI Magazine 13(2), pp. 81–95.

2. Hayes, P. and Ford, K. (1995), 'Turing Test Considered Harmful', in Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Vol. 1, pp. 972–977.

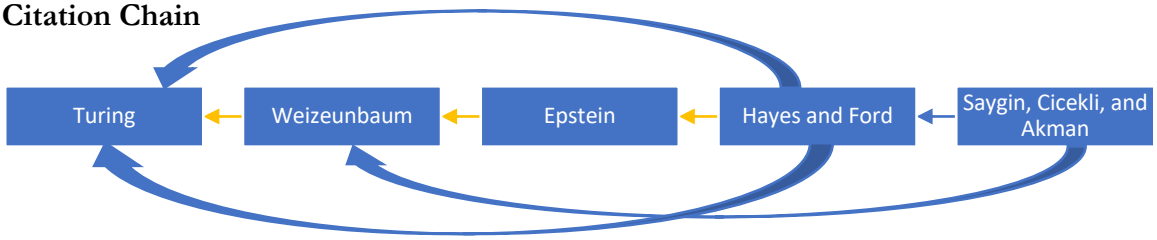
3. Levesque, Hector J, Davis, Ernest, and Morgenstern, Leora. The winograd schema challenge. In AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, 2011.

4. Saygin, A., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines, 10*(4), 463–518. https://crl.ucsd.edu/~saygin/papers/MMTT.pdf.

5. Turing, A. (1950), 'Computing Machinery and Intelligence', Mind 59(236), pp. 433–460.

6. Weizenbaum, J. (1966), 'ELIZA–A Computer Program for the Study of Natural Language Communication Between Men and Machines', Communications of the ACM 9, pp. 36–45.

Citation Chain



Note: Orange line indicates that the paper that does not directly cite the other paper, but mentions the paper's topic directly by name. Weizenbaum mentions the "turing test," Epstein mentions "ELIZA," and Hayes and Ford mention the "Loebner competition."