JORDAN BOYD-GRABER

QUESTIONING ARTIFICIAL INTELLIGENCE

UNIVERSITY OF MARYLAND

3

The Turing Test: The Deceptions that Defined Artificial Intelligence

In the 1950s, Alan Turing proposed a parlor game that would come to define the artificial intelligence: could a wiley interrogator discern whom they were talking to just through posing clever questions. The eponymous Turing Test is the most durable (but contentious) definition of what an intelligent computer is, but it is built on deception.

Topics in chapter:

1. **Turing's Legacy.** Why Turing is widely considered the father of computer science for multiple reasons and how contributed to cloak and dagger spy operations.
2. **The Imitation Game.** A gay man in England—where homosexuality is criminalized—posits a games about the ability of someone to pass in different gender roles. Why this is difficult lays the foundation for why computers still struggle to capture human nuance.
3. **No, the Turing Test has not been Solved.** While many have claimed to pass the Turing Test, it hasn't happened yet. We review some notable claims of passing the Turing Test and why they fell short, and why you shouldn't fall for the claims of charlatans.
4. **A Rigorous Test.** The ingredients that make a Turing Test legitimate
5. **General Artificial Intelligence.** How the holy grail of AI is defined and how the Turing Test can measure it.

The history of artificial intelligence begins in many ways with a question answering game called the Turing Test. This idea came out of the research of a researcher at the University of Manchester named Alan Turing. But the history is much deeper than that; there are spies, misrepresentations, deceptions, and attempting to unmask pretenders. This chapter shows how this history of deception defined how we approach AI today.

3.1 Turing's Legacy: A Cloak and Dagger Computer Problem

If you haven't heard of Alan Turing for his theoretical contributions to computer science, you may know him as one of the scientists who decoded the Nazi

Enigma device (which is why his memorial bench has ciphertext underneath his name, Figure 3.1).

Figure 3.1: A memorial bench of Alan Turing in Manchester, England. The ciphertext under his name represents his contributions to breaking the Enigma code.

While now the story is well known, at the time it was one of the world's most closely kept secrets. To prevent the Germans from knowing that Turing (and the rest of the Bletchley Park team) had cracked the code, the British fabricated alternate explanations of how they gathered valuable intelligence. At one level they attributed the intelligence to a British-run spy ring in German called 'Boniface', and then was called 'Ultra' (Gilbert, 2015). For example the British engineered a flyover by a reconnaissance plane to "spot" a German ship as an alternate explanation for how the British knew to attack them (Wilcox and for Cryptologic History., 2001).

Turing's life is incredible. Beyond saving thousands of lives by breaking German codes, Turing hypothesized what a computer could do. His abstraction of a Turing machine is the field's definition of what it means to know what a computer can do. If a problem can be solved on a theoretical Turing machine, we call it "computable"; any computer made using standard logic[1] can also solve the problem.[2]

[1] E.g., this applies to any computer that Intel will put out but not to quantum computers (should they ever work).

[2] Assuming you have sufficient memory and time.

3.2 The Imitation Game

In addition to the deception that the British government did to hide Turing's work, Turing himself had to deceive the British government about his own life. To put it bluntly, Turing was a gay man living in a country that criminalized homosexuality. He had to hide his true self and pretend to be someone else.

In 2009, the Prime Minister of Britain issued a formal apology (Brown, 2009):

> It is no exaggeration to say that, without his outstanding contribution, the history of World War Two could well have been very different. He truly was one of those individuals we can point to whose unique contribution helped to turn the tide of war. The debt of gratitude he is owed makes it all the more horrifying,

> therefore, that he was treated so inhumanely. In 1952, he was convicted of 'gross indecency' - in effect, tried for being gay. His sentence - and he was faced with the miserable choice of this or prison - was chemical castration by a series of injections of female hormones. He took his own life just two years later.

Whole books (and films) have been devoted to his codebreaking and his bittersweet personal life. But Turing is also a game designer. Quoting from Bishop's description:

> Turing called for a human interrogator (C) to hold a conversation with a male and female respondent (A and B) with whom the interrogator could communicate only indirectly by typewritten text. The object of this game was for the interrogator to correctly identify the gender of the players (A and B) purely as a result of such textual interactions

This too is a deceptive exercise: the man must lie to win the game. The key to correctly deciding the genders of the players is more about determining which player lacks key knowledge about the experience of being a man or a woman.[3]

This is hard because there are lots of "tells" that can expose someone who is transgressing against their assigned gender roles. A skilled interrogator can ask about how it feels to have a period, the rules for picking which urinal to use in a crowded restroom, or how to respond to a girlfriend asking about her outfit.

And while an impostor can read many accounts to get an idea of what someone has said about these topics, simply parroting these stories does not equate to understanding. This is analogous to current AI craze: they have read all of the Internet, and the challenge is to measure whether all of that ingested knowledge can fuel an incipient intelligence.

[3] What makes the Turing Test more poignant is that in a country where homosexuality is illegal, every interaction of Turing is dangerous: Turing is pretending to be something that he is not but has to present himself as if he were a heterosexual man. Every conversation is a possible interrogation to see if he can "pass". There's no evidence evidence to suggest that this was an inspiration for his first Turing Test, however.

3.3 The Second Turing Test

With that being said, lets turn this back to AI. Turing then thought, *let's replace the man with a computer*. In other words, the computer is doing the deception. If the interrogator cannot determine whether the entity on the other end of the teletype is a human or a machine (even when they're both claiming to be a machine), then the machine has displayed something that looks like intelligence.

Why do we need this silly game? Because if you picked something that correlates with intelligence (say playing chess), you could fairly easily make a machine that can only do that. Turing believed—and enough have gone along with this—that answering questions about any topic is a general enough test of intelligence that it would require something of equivalent intelligence to a human to win the game.

To recap, to have a Turing test, you need **A Skilled Interrogator**, **An Unrestricted Domain**, and **Knowledge that You're Playing the Game**. As

we'll see in a second, many of the times have claimed that the Turing Test has been past fail because one of these ingredients has been lacking.

No, the Turing Test has not been Solved

But I want to emphasize not just the broad strokes of the Turing Test, but why we need to think critically about things that call themselves the Turing test. Also, it's more fun than talking abstractly about the Turing test.

Let's start with PARRY, a system designed by Kenneth Mark Colby to simulate a paranoid schizophrenic (Colby, 1981). And when you connected psychiatrists to either real patients or PARRY, they couldn't tell the difference. So here the problem is the judges. Not because they aren't experts–they are–but because their backgrounds prevent them from being effective judges.

Psychiatrists are doctors bound by the hippocratic oath: they cannot ask probing, in-depth questions that might harm a patient. So, thus, this really isn't a Turing test. The interrogators weren't unchained to truly ask whatever questions they needed to ask.

Stu Schieber has a great take on this problem; I'd encourage you to read the whole thing. Rather, instead of a "take", it's more of a "takedown" of a competition called the Loebner prize that purports to implement the Turing Test. This was a part of a more general trend of over-promising what AI can do but under-delivering even in 1991. One of the conditions of the Loebner prise was that the judges could not ask tricky questions or use guile and significantly limited the time judges could interact with a program.

Taken together, Shieber (1994) contends

> Thus, it is difficult to imagine a clear scientific goal that the Loebner prize might satisfy. Turing's test as originally defined, on the other hand had a clear goal, to serve as a sufficient condition for demonstrating that a human artifact exhibited intelligent behavior. Even this goal is lost in the Loebner prize competition. By limiting the test, it no longer serves its original purpose (and arguably no purpose at all), as Turing's syllogism fails. It is questionable whether the notion of a Turing test limited in the ways specified by the Loebner prize committee is even a coherent one.

Another example that some people claim is an example of AI passing the Turing Test is Google Duplex: Google offers to call a restaurant to make a reservation for you. Their text-to-speech system is very good, but also puts in disfluencies such as adding "um" and pauses to make it sound more human. Here, the judge—a poor service worker taking a reservation and getting paid minimum wage—is fooled into thinking that they're talking to a human. But this doesn't count either because the judge doesn't know they're a judge! Part of the social contract (at least for now) is that a recipient of a phone call assumes that they are getting a call from a human until proven wrong.

All of this doesn't mean that the Turing Test is flawed. It has remained a part of AI for three quarters of a century because it's a simple, intuitive test of whether we have achieved artificial intelligence. So although we haven't

had a real Turing Test yet, a judge asking questions of either an AI or a human remains many researchers' goal.

Criticisms of the Turing Test

But it's worth pausing to consider some of the notable criticisms of the Turing Test. One of the best known criticisms of the Turing Test are inspired by Searle's Chinese Room: imagine a monolingual American from Omaha trapped in a box with an incredible collection of translations. Chinese text comes in, and he needs to provide a Russian translation. He understands neither language, but can look up the appropriate translation from his collection of books. To an outside observer, he has full mastery of both languages. Does this Omahan actually possess that understanding?

There are several counterarguments to this criticism: you're not evaluating the man himself but rather the system he is a part of. The room, plus the astounding translation materials (and whoever put them together) have created a system that can perform perfect translations. While I think there is some merit to this defense, I prefer the "interactive proof" argument by Stu Shieber (Shieber, 2007).

Sheiber argues that the number of sentences in any language is so unfathomably large that the Turing Test is able to probe whether the effect is mere memorization or not. With reasonable assumptions, a Turing test of, say 300 turns would have a vanishingly small probability of a false positive. In other words, it would be impossible to encode everything necessary in the Omahan's lookup to deceive a skilled inquisitor.

3.4 The First Turing Test: Deception and Theory of Mind

Sterrett (2000) argues that Turing actually proposed two versions of his eponymous test (Table 3.1).[4] There are actually two ways of interpreting "replace the man with a computer" in the Imitation Game. A more literal interpretation would be a computer and a man would both compete who could best impersonate a woman.

Sterrett argues that this is a more fair test because it puts the human and the man in a more similar role: they are both pretending to be something they are not. Sterrett (2000) says that a dullard in the Second Turing test must do nothing more than answer questions honestly; there's no skill involved. As we argue in Chapter 10, we are often interested in comparing human vs. computer skill. This scenario has humans and computers both compete to pretend to be a woman, and you could measure the relative probability of success.[5]

What I like about this opportunity for comparison is that it requires humans and computers to explicitly use theory of mind(): they must create a fictitious persona that will be probed by the interogator and also craft responses in a

[4] This is an argument because Turing's specification is vague; there is a question whether the ambiguity was inadvertent or intentional. I prefer to think of it as intentional, but it fits with this chapter's theme of connecting the Turing Test with deception, so take my self-serving interpretation with a grain of salt.

[5] While I am also a fan of this interpretation of the Turing Test, it is less well-defined and standardized. An obvious criticism is that you would probably want the interrogator to always be a woman.

	Original Imitation Game	First Turing Test	Second Turing Test
Interrogator	Identify which hidden participant is really the woman, based only on written replies.	Identify which hidden participant is really the woman, based only on written replies.	Determine which hidden participant is human and which is a machine, based only on conversation.
Control	Answer truthfully as the woman, helping the interrogator make the correct identification.	Answer truthfully as the woman, serving as the baseline for successful identification.	Respond naturally as a human, making it possible for the interrogator to recognize them as human.
Candidate	A man impersonates the woman well enough that the interrogator cannot reliably distinguish it from the human impersonator.	A man **or a computer** impersonates the woman well enough that the interrogator cannot reliably distinguish it from the human impersonator.	Respond in a way that is indistinguishable from the human participant, so that the interrogator cannot reliably tell it is a machine.

Table 3.1: Different versions of the imitation game and the Turing Test, distinguished by participant roles and goals. While the original Imitation Game was about a man impersonating a woman, the Second Turing Test is conventionally what people mean when people talk about *The* Turing Test. However, I agree with Sterrett (2000)'s take that the **First Turing Test** is implied by Turing's proposal, and in many ways make more sense.

way that will elicit the correct responses from the interrogator.[6]

But why should we put deception on such a high pedestal? A cynic would that all humans lie, so we we want to have an AI that can do cognitively difficult tasks, it should also be able to lie as well as a human. (Henricksen, 2024) argues that children learning to lie is not a good outcome, lying shows that the have mastered important skills: knowing what others know (sometimes called "theory of mind"), navigating social interactions ("self control"), and being able to convey the veracity of your statements to others ("presentation"). Being able to sustain a lie is more difficult than simply reporting the truth.

And this is what makes Sterrett's argument attractive from a measurement perspective. When a computer impersonating to be a woman is pitted against a man attempting to answer as a woman, we can look at their relative success rates. Thus, this test would be better able to measure incremental progress.

An example of a modern grand challenge that *felt like* a Turing Test—but ultimately was not—was Meta's claim that CICERO had mastered the game of *Diplomacy*, a framing that you see echoed in the popular coverage surrounding its publication in *Science*.[7]

Diplomacy is a seven-player game set at the eve of Word War I, and each player takes the role of one of the great powers of Europe. The goal is to take over the entire map through skillfully moving fleets and armies across the map. It's hard to best the description of the game that that appeared in the

[6] If asking questions about gender is offputting, you could easily replace this with an American pretending to be a Brit; the same principles apply.

[7] If you look at the article on *Science*, the headline itself is "AI masters Diplomacy". While that is from the editor of the article, not the authors of the paper, it reinforces this false narrative of overselling AI.

Times obituary of the game's creator, Allan Calhamer (Fox, 2013):

> In each of the game's compulsory negotiation periods, which involve whispering furtively in corners while simultaneously routing eavesdroppers, players in weaker positions band together against those in stronger ones.
>
> What emerges from these sessions, which govern the moves on the board, is a world of quicksilver alliances: joint military campaigns are planned; deals are made, then abrogated, and new agreements arise to take their place. Foe is friend and friend is foe, and it is seldom possible to tell the two apart.

Unlike chess, the difficulty is not calculation alone but persuasion, trust, and betrayal. In that sense, the game—like the Turing Test—is built around deception: success depends on anticipating what others believe and sometimes saying something untrue or misleading to get another player of the game to do something against their interest. For example, Italy might might offer England an alliance: let's work together to take out France together. But England's move might leave them open to an attack by agreeing to the plan, allowing Italy to benefit from a bait and switch.

It is therefore unsurprising that CICERO was widely interpreted as a kind of Turing Test milestone. Meta reported that human opponents rarely suspected they were playing against a machine, and online commentary quickly jumped to "passing the Turing Test".

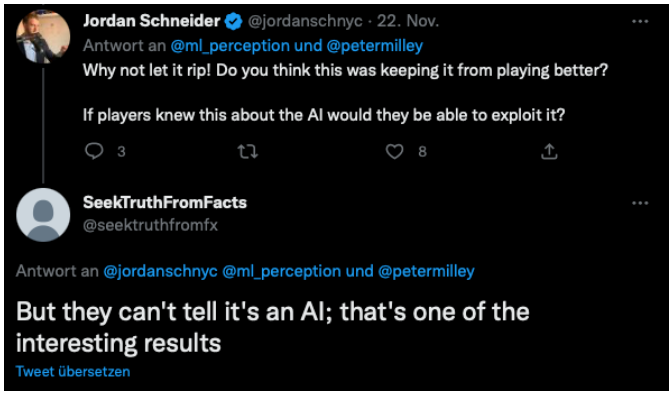

Figure 3.2: After CICERO was unveiled, the reporting focused on how players could not detect that they were playing against a computer, leading to comparisons with the Turing Test.

As with IBM Watson's victory on *Jeopardy!* (Chapter 6), the temptation was to treat success in a highly visible game as evidence that the broader problem of intelligence had been solved. At the risk of spoiling some of that argument, there's more to *Jeopardy!* than answering questions, muddying the scientific question of whether Watson is smarter than humans.

Taking a step back, suppose that you built a baseball playing robot that was preternaturally good at hitting dingers out of the park. But couldn't catch a ball to save its electronic life and can only move at two miles an hour around the bases (I imagine it beeping constantly as it does so). Oh, and it can pitch, but it only sends balls straight down the middle. It ends up winning a bunch of games because it's indefatigable and can outscore the weak humans. Does

that mean that it has mastered baseball? I think most people would argue that no, there's more to baseball than hitting dingers: you need to run, catch, throw a ball, pitch curveballs and sliders.

It's the same with Diplomacy: to master Diplomacy, you need to be able to detect lies and imminent betrayal, from other players. And I'd also say as a sometime amateur player of Diplomacy, you need to execute betrayal as well. This was also the conclusion of the excellent Economist write up of CICERO, although I disagree with its assessment that another game "has fallen" (The Economist, 2022):

> Seasoned Diplomacy players will, though, want to know something else: has it learned how to stab? Stabbing—saying one thing and doing another (especially, attacking a current ally) is seen by many as Diplomacy's defining feature. But, though Cicero did "strategically withhold information from players in gameplay", it did not actually stab any of its opponents. Perhaps it was this final lack of Machiavellian ruthlessness which explains why it was only in the top 10%, and not victor ludorum.

The conditions of the experiment departed sharply from the spirit of a Turing-style interrogation. Players did not enter the game as skilled judges trying to unmask a computer, conversations were constrained by blitz settings, and no one was told that a machine might be among them. Fooling an unsuspecting opponent is not the same as surviving questioning by an interrogator who knows that deception is the point.

As is common with understanding AI claims, if you focus the evaluation on what you care about, a different story emerges. Thus, my student Joy Wongkamjan tried to answer was CICERO a persuasive and deceptive negotiator, or because it was simply an exceptionally strong tactical player? Across two dozen games and more than 200 hours of human play, they find that Cicero does indeed win against human opponents—often decisively—but that its language falls short of the social intelligence implied by popular coverage: other humans can reliably detect CICERO (Figure 3.3), CICERO is less persuasive than humans, and humans players trust CICERO less than humans even though CICERO lies less than humans.

The lesson is not that CICERO was uninteresting—it was a remarkable technical achievement—but that we should be cautious about what such victories actually demonstrate. A true test of social intelligence would need to isolate communication itself: the ability to form commitments, detect betrayal, persuade others to change course, and maintain coherent intentions over long interaction. Without that, claims of "mastery" risk mistaking fluent dialogue and strong tactics for the deeper competencies that the Turing Test was meant to probe.

3.5 *A Rigorous Test*

So let's be true to the spirit of Turing's idea of a parlor game. Let's make it visible to the public, let's refine the rules and the judges to make it more realistic and more fun. By putting these games in the public view and letting judges learn the best strategies for discerning humans from computers, both sides can become worthier adversaries.

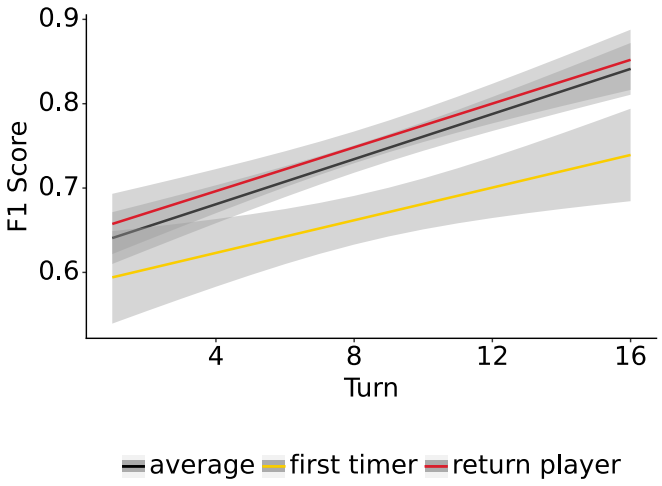

Figure 3.3: Although Meta claimed that human players did not recognize when they were playing against computer opponents (an unannounced, impromptеau Turing Test), **?** show that players in a game of *Diplomacy can* reliably detect an AI opponent. F_1 Score is a combination of precision and recall, as introduced in Section 4.1 (Equation 4.3).

And although few people are running Turing Test experiments today, it helps motivate the kind of skills we want to see in our AIs and what it would take to scientifically measure those skills. And from a public interest perspective, I think putting this slow advance of ever more capable computers answering trickier questions should be out in front of the public. Not just to keep the interrogators honest but to also keep the companies and interests that sell AI honest, as deception is not just a part of the Turing Test, it's an unfortunate part of our increasingly AI-infused reality (Chapter 13). The public should know not just how to avoid AI scams but to know the limits of AI, and this is a fine way to make that public. But on the other side of the coin, it is also worthwhile for the public to know when AI has really advance…the public has a right to know how computers react to challenging scenarios. Better to see them first played out for fun in a game than in high-stakes transactions, a doctor's office, or a courtroom.

But above all, this only works if we have good questions, so if we believe that the Turing test really is the holy grail of AI, we as humans need to know how to ask the right questions and computers need to be able to answer any question that's thrown at them. The next chapter shows the earliest attempts to get computers to answer questions only decades after Turing first proposed using deception to test machine intelligence.