Class Ten: The Turing Test

Philosophy and Science Fiction - Ryan Simonelli

October 31, 2022

1 The Question of Consciousness and AI in *Ex Machina*

- **Caleb as Turing Tester:** In *Ex Machina*, Caleb is brought in to Nathan's remote compound to perform the "Turing Test" on Ava, a new AI system that Nathan has developed.
  - **What he's testing for:** It's not made entirely clear what, exactly, Caleb is supposed to be testing for, but it seems like we can distinguish between two basic things:
    - *Intelligence*, both cognitive and emotional.
    - *Consciousness*, whatever that is supposed to mean.

    It's not entirely clear how these two things are related, but, clearly, they are in some way.
  - **A More Helpful Distinction:** Recall the distinction between *sentience* and *sapience*:
    - *Sentience*, general sensory awareness, both of one's external environment and of internal states of oneself.
    - *Sapience*, the capacity to reflect and think rationally about things, intimately tied with the ability to speak a language.

    A mouse is sentient, and, in that sense, we'd say it's conscious. But it doesn't seem like it's conscious in the sense that Caleb's testing Ava for (if he established that the mind of Ava was comparable to that of a mouse, he wouldn't say she passes the test). So it seems that the kind of consciousness that he's testing for has more to do with *sapience* rather than mere *sentience*.
    - *Note:* These terms often get mixed up in sci-fi and popular culture more generally, but this is the way they're generally used in philosophy.

    Arguably (see Matt Boyle's work), sapience isn't something that's simply *tacked on* to sentience, leaving it as is, but, rather *transforms* what it is to be sentient.
- **Caleb as Part of the Test:** Towards the end of the movie, we realize that the *real* test includes Caleb as a part. He is not really the tester, but an element of the test. The real test, as Nathan conceives of it, is to see if Ava can convince Nathan she's conscious so as to use him as a means of escape.
- **Ava Passes with Flying Colors:** Not only does Eva convince Caleb that she's intelligent and conscious, making him feel empathy towards her, but she *plays him*, using these feelings to her advantage to escape by herself from the compound.
- **Question:** Given that she passes the test, Ava's behavior clearly seems to be indicative of intelligence and consciousness. Can we conclude from this that she *really is* intelligent and conscious?

2 The Turing Test

- **Alan Turing:** Logician, mathematician, philosopher, and cryptographer extraordinaire.
  - One of the founders of the field of computer science, along with his doctoral adviser Alonzo Church.

  - Arguably the sole founder of the field of artificial intelligence.
  - Helped crack Nazi codes in World War II.
  - Played by Benedict Cumberbatch in the 2014 movie *The Imitation Game*.
  - Persecuted for homosexuality and likely killed himself (in 1954, at the age of 41) because of it.

- **The Imitation Game:** Three parties:
  - **A:** A participant of the *target type* (a woman, a human, etc.).
  - **B:** A participant *imitating* that type (a man, a machine, etc.).
  - **C:** An *interrogator* or *tester*, who must guess who is the member of the target type.

  B *passes* the test if the interrogator is no better than chance at determining who is target type and who is the imitator.

- **Turing's Proposal:** The way to answer the question of whether machines can think is consider whether they're able to pass the imitation game, passing as a human being.
  - **Question:** In the original example of the imitation game, where there is a man pretending to be a woman, if that person passes the test, we're not inclined to say that this person *actually is* a woman, just that they're capable of imitating one. Why should we think any different about a machine imitating conscious thinking?

- **The Basic Thought Underlying the Proposal:** What matters in determining intelligence is not the specific hardware or programming but, rather, *behavior*, and, specifically, the radically open behavior of engaging in conversation about any topic whatsoever.
  - Contrast a computer that just plays chess, passing a chess only Turing test. Clearly, we're not inclined to say, on that basis, that this computer is conscious/intelligent. What distinguishes this from the actual Turing test is the *breadth of capacities* needed to pass the latter but not the former.
  - **A Further Thought:** If something really did pass the test (at least on some variants of it, which we'll consider), we *couldn't help but believe* it was conscious, just as we can't help but believe our fellow humans are conscious, and so we might as well conclude that such a thing really is conscious?

3 Varieties of Turing Tests, Takers, and Questions

- **The Mere Turing Test, and Tricky Turing Test Takers:** If the goal is simply to pass as any human, without any restriction on what sort of human one needs to be, a test (with a relatively short time scale) can be easily passed by programs imitating people who, for some reason or another, aren't willing or able to engage in a normal lucid conversation.
  - **PARRY:** In 1972, a program called "PARRY" passed the Turing test by simulated the behavior of a paranoid schizophrenic.
  - **Eugene Goostman:** In 2014, a program called "Eugene Goostman," which simulated a 13-year-old Ukrainian boy, passed the Turing Test.

  Clearly this is not what Turing had in mind.

- **The Long Turing Test:** Just like the regular Turing Test, but played out, say, for two hours a day over the course of a year, so that the interrogator has the potential to develop a deep interpersonal understanding with the participants that builds over time.
  - **Question:** Imagine the original example of the Imitation Game, with the man and woman roles, played out for

- **The Honest Turing Test:** One can consider a "variant" on the Turing Test (though it's not really clear whether to still call it a "Turing Test" at this point) in which one knows that one is speaking to an AI, but still evaluates whether its responses are sufficiently human-like to attribute to it intelligence.
  - This is the variant of the test considered in Ex Machina.

4 Different Kinds of AI

- **The Original Notion of a "Computer":** Back in the early 1900s and before, the term "computer" mainly applied to *people* who would *compute* things, performing calculations.
    - This is what you did back in grade school math class, for instance, when you did long division.
        - You had an "algorithm," a set of rules for doing the computation, you got a pair of numbers, and you applied the algorithm to these two numbers, dividing one by the other.
- **"Good Old Fashioned AI":** One can think of a simple computer program in terms of a set of rules for turning inputs into outputs, where this is essentially a sort of symbol manipulation.
- **Simple Conversation Programs:** You can learn in an hour or so how to write a program that simulates a conversation like the following:

    AI: Hello, what is your name?
    ME: Ryan.
    AI: Hello, Ryan! How old are you?
    ME: 30.
    AI: Wow! That's old! I'm only 2 years old, so you're 28 years older than me!

    It's quite obvious how one could write a program that does this, but, on the most obvious way of doing this, one will get defective conversations like the following:

    AI: Hello, what is your name?
    ME: My name's Ryan.
    AI: Hello, My name's Ryan! How old are you?
    ME: No, my name is "Ryan," not "My name's Ryan."
    AI: What was that? Can you tell me your age, as a number?
    ME: 30.
    AI: Wow! That's old! I'm only 2 years old, so you're 28 years older than me!

    Of course, this is a hopelessly simple program and it could not possibly pass as thinking, but early AI programs, such as Joseph Weizenbaum's 1966 "ELIZA" (which you can talk with here: http://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm) are see-through in basically the same way.
- **Learning AI:** In the last section of "Computing Machinery and Intelligence," Turing predicts the advent of "Learning Machines," artificial intelligences which, rather than being programmed to be intelligent from the outset, are programmed to be capable of learning and are then subjected to an education process.
    - We'll talk more about the details of such AI constructions next class.

5 Objections to Machines Thinking

- **The Theological Objection:** Thinking is a god-given capacity that only human beings can have because only human beings have a soul.
    - **The Basic Response:** Theological objections like this often seem good at a time, but are almost always lose their force in the course of time (consider, for instance, such objections to abiogenesis).
- **The Mathematical Objection (developed later by J.L. Lucas):** A machine could not possibly represent its own formal system so as to be able to answer a question that makes reference to that very system. It seems, however, that we always "jump outside" of any formal system, and so our own minds cannot possibly be any such system.
    - **The Basic Response:** All this shows is that we are more capable of abstraction than any simple machine that we've created so far—it doesn't show that there are no bounds on our intellect that only a machine smarter than us could grasp.

- **The Consciousness Objection:** Machines might be able to do lots of things—for instance, write a sonnett—but until they are capable of doing these things for the sorts of conscious reasons that we do them, they should not be regarded as intelligent.
    - **The Basic Response:** The test enables one to do more than simply ask for the participant to produce a sonnett—one can also ask about the reasons one wrote the sonnett that one did. If one denies that responses to such queries count, then it seems like one undercuts any evidence one might have that other people are conscious.
- **The Argument from Various Lacks of Abilities:** Machines will never be able to do X, where X is something thought to be distinctively human.
    - **The Basic Response:** No direct evidence is ever given for these claims, and some of them (for instance, that machines can't make mistakes) are just confused.
- **Lady Lovelace's Objection:** Machines can only follow a pre-set set of instructions. They can't do anything new.
    - **The Basic Response:** Even purely algorithmic machines can suprise us, since we might make mistakes that they can catch (consider a spell-checking machine) or they might extract consequences that we do not ourselves extract. More importantly, consider learning machines.
- **The Argument from the Informality of Behavior:** Our behavior is not governed by any definite set of rules for conduct.
    - **The Basic Response:** This objection hinges on the ambiguity between two senses in which behavior may be "governed by rules," acting *under a conception* of a rule and acting *in accord* with a rule. Clearly, our behavior is not governed by a definite set of rules in the first sense, but the AI's doesn't need to be either, and it's not clear our behavior is not governed by a definite set of rules in the second sense.

6 Further Questions We Have

- **The Ethical One:** Turing's discussion explicitly revolves around the question of whether we *can* construct intelligent or conscious machines; it ignores the question of whether we *should*. In the movie the latter question is highlighted just as much as the former. Supposing it's *possible* to construct genuine artificial intelligence, should we *actually do so?*