Thinking and Autonomy

Prasad Tadepalli

School of Electrical Engineering and Computer Science
Oregon State University
Turing Test (1950)

- The interrogator C needs to decide if he is talking to a computer B or a human A.
- If he is not able to distinguish A from B in ‘a reasonable time,’ then either
  - Both A and B think, or
  - Neither of them think
- It is more polite to say that both of them think
Arguments against Turing Test

- Mathematical Objection: There are many problems for which there are no algorithms, e.g., the halting problem. So machines cannot solve them.
  - True, but there is no evidence that people can solve them either.
- Lack of Originality (Ada Lovelace): A machine can’t do anything original because it needs to be programmed
  - Machines can *learn* to do things that surprise their programmers
  - The interrogator can ask questions that require creativity and conduct quizzes for originality
    - “Can you write a poem?”
    - In your poem you say, “Shall I compare thee to Summer’s night?” Why not Spring night? Why not Winter’s night?
- Lack of consciousness (Prof. Jefferson, Prof. John Searle)
Searle’s Chinese Room

- Imagine a native English speaker that does not know a word of Chinese locked in a room in China.
- He has a rule book that allows him to perfectly answer any question put in Chinese by manipulating Chinese characters.
- He is so fast and so good in following the rules that he answers every question that is put to him as well as a native Chinese speaker.
- Does he understand Chinese?
Searle’s Argument in Slow Motion

- Turing is equating input-output behavior with thinking (this is called “strong AI” by Searle, “functionalism” in philosophy)
- In principle, the same behavior can be caused by multiple means
- It is most likely, and at least conceivable, that machines exhibit intelligent behavior in a very different way than humans
  - Chess and Go machines search millions of positions. The neurons are too slow for people to be able to do this.
- Hence behavior alone is not a sufficient indicator of thinking.
- Searle considers many counterarguments to his thought experiment
  - System’s reply: the person may not understand, but the room does
  - Robot reply: if you replace the Chinese room with a robot that manipulates the real world (rather than symbols), the robot is indistinguishable from a person who understands Chinese
  - Pragmatic reply: It is not practical to follow the rules as quickly as a Chinese speaker would
Why does it matter?

- South Korea’s Super Aegis II has a supergun with a range of 4 kilometers, and is powerful enough to stop a truck.
- Costing $40M, it was deployed in Korea’s DMZ, and sold to many Arab countries.
- Safety features:
  - It is required to issue a warning before shooting. Has a voice that can reach 3 kilometers.
  - The shooting requires an OK and a password from a human operator.
  - Safeguards are largely self-imposed.
- Are these safeguards sufficient? Are we entering a new era of “killer robots”? 
AI and Ethics

- Trust
  - Will you trust a robot to babysit your child or care for your parent?
  - Can we trust the self-driving cars?
  - Is behavioral testing sufficient to gain trust?

- Ethics
  - Will you let AI make a life-saving medical decision?
  - Defend a sensitive border?
  - Is it ethical to (not) let machines make critical decisions?

- Pragmatics
  - Can one recognize a bad decision before it is too late?
  - Who is responsible if something goes wrong?

- Social status of robots
  - Can/should machines ever make moral judgments?
  - Is it OK to have a robot for a friend?
  - Should robots have rights?