Lecture 2
The Turing Test and Searle's Chinese Room

Rob Gaizauskas

Lecture Outline

- How can we decide if a machine is intelligent?
  - Turing Test
  - Arguments Against the Turing Test
- Is AI Possible?
  - Searle's Chinese Room Argument
  - Arguments Against Searle

How can we decide if a machine is intelligent?

- Suppose someone presents us with a machine they claim is a ``thinking'' machine or is ``intelligent''.
- What test(s) should we use to determine is this is true?
- The most famous test is the Turing Test, named after the British scientist Alan Turing, who proposed it in 1950.
  - Only 4 electronic computers in existence
  - Before Dartmouth conference (1956) and the `birth' of AI

Turing Test

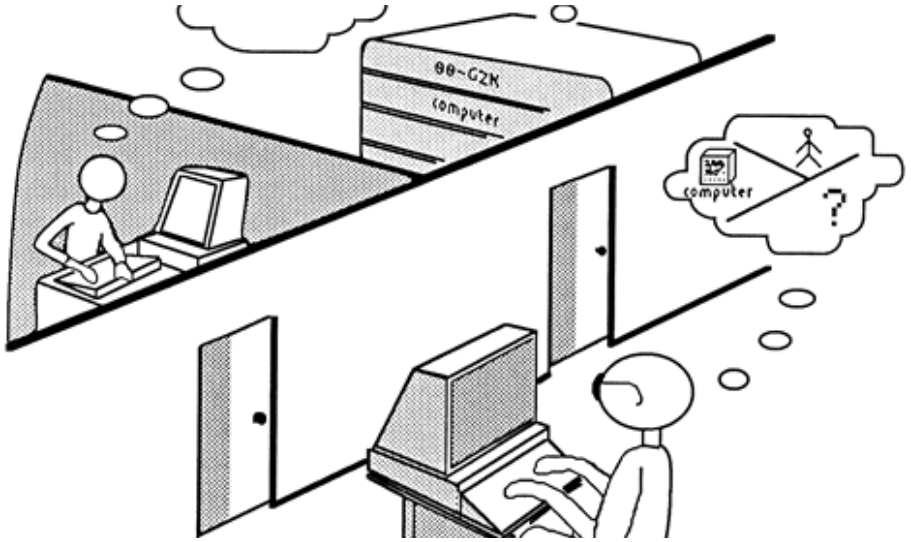
- In his paper "Computing Machinery and Intelligence", Turing proposed to replace the question "Can machines think?" with a question framed in terms of the imitation game
- In the imitation game there are three players:
  - A man (A)
  - A woman (B)
  - An interrogator of either gender (C)
- C, the interrogator, is in a separate room and communicates with A and B, who he knows as X and Y, only by asking written (typed) questions or via an intermediary
- Object of the game
  - For C is to determine which of X and Y is A and B, i.e. which is the man and which the woman
  - For A is to mislead C into making the wrong identification, i.e. convince C he is the woman
  - For B is to help C make the correct identification

Turing Test

- Turing's twist is to suggest A (the one trying to deceive) be replaced by a computer
- The question "Can machines think" is now replaced by the question:

  "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman ?"

  (Turing , 1950)



From: www.alanturing.net

Turing's Prediction

- "I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning."
- It's now 2015 – 65 years after Turing made this prediction.
- Are we there? – no – why not? – more on this in a later lecture

“Eugene Goostman”

- Chatterbot developed in St Petersburg in 2001 by 3 Russian/Ukrainian programmers
- Pretends to be a 13 year old Ukrainian boy with English as a 2nd language
  - "not too old to know everything and not too young to know nothing”

- On June 7, 2014 (60th anniversary of Turing’s death) Goostman was claimed to have passed the Turing test
  - Event at Royal Society, London, organized by Kevin Warwick of Reading University
  - 1/3 of (30) judges fooled after 5 minute conversation
- Reaction from AI community mostly negative, arguing
  - Chatterbot used its age/language limitations to trick judges
  - Turing’s “threshold” meant as a prediction about where research would be in 50 years, not as a sufficient condition for deeming a computer intelligent
  - Devalues “serious” AI research – nothing new in the approach

The Loebner Competition

- Loebner competition runs annually
  - $100,000 prize for first computer to fool the judges
  - Annual prize ($4000 in 2015) for most convincing (="least bad") computer
  - No computer has yet fooled the judges
- In 2015 run at Bletchley Park (September 19th)
- Video chat with the winner: 2013 Loebner Prize Winner
- More details at:

  http://www.loebner.net/Prizef/loebner-prize.html

  and

  http://www.aisb.org.uk/events/loebner-prize

Class Discussion

- Is the Turing Test a good test for thinking/ intelligence?
  - Is it too easy or too difficult?
  - How could it be improved?

Objections Anticipated by Turing

Turing anticipated 9 objections to his position:

1. The theological objection *
2. The 'heads in the sand' objection*
3. The mathematical objection
4. The argument from consciousness *
5. Arguments from various disabilities
6. Lady Lovelace's objection *
7. Argument from continuity in the nervous system
8. The argument from informality of behaviour
9. The argument from extra-sensory perception

The theological objection

"Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think"

- Turing's dismissal: Why not believe that God could give a soul to a machine if he wished?
- He adds:

  "I should find the argument more convincing if animals were classed with men, for there is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals."

The heads in the sand objection

"The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so."

- Turing's dismissal:
  - "I do not think that this argument is sufficiently substantial to require refutation. Consolation would be more appropriate"
  - objection related to theological objection in that people like to believe they are somehow superior to the rest of creation

Argument from consciousness

"This argument is very well expressed in Professor Jefferson's Lister Oration for 1949, from which I quote. 'Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain – that is not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants' "

Turing's dismissal of argument from consciousness

- How could we tell that a machine that had passed the test was not conscious?
- The argument seems to suggest the only way we could be sure that a machine thinks is to be that machine and to feel oneself thinking.
- But, similarly, the only way to be sure that someone else thinks is to be that person.
- How do we know that anyone else is conscious/thinks? – leads to philosophical position known as Solipsism.
- But we assume other people are conscious. Similarly we could assume that a machine that passes the Turing test is effectively conscious.

Lady Lovelace's objection

- Charles Babbage (1792-1871)
  - Specified a general purpose calculator called the Analytical Engine
  - Was in fact Turing-complete (i.e. could compute every function a general purpose modern computer could)
  - entirely mechanical, and never actually built.
- In her 1842 memoir of Babbage Lady Lovelace wrote: "*The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform*" (her emphasis)
- Turing suggests variants of this objection
  - a machine can ' never do anything really new '
  - a machine can never ' take us by surprise '
- A computer cannot be creative, it cannot originate anything

Turing's dismissal of Lady Lovelace's objection

- Points out that all creative acts could be seen as following from "seeds" planted in us by teaching, i.e. that humans never really do anything new either
- Further, computers can and do surprise their programmers – producing answers that were not expected.
  - Data may have originally be given to the computer, but then it may have been able to work out its consequences and implications
  - Consequences may include much not foreseen by the programmer

Criticisms of Turing's Position

Copeland (1993) raises 4 objections to the Turing Test

1. The chimpanzee objection:
   - Chimps can think, but could not pass the Turing Test (same for dolphins, dogs, pre-linguistic humans). Therefore, test is too strong.
   - Follows that Test cannot replace (i.e. exactly substitute for) the question "Can a computer think?"
   - But, while not a necessary condition for thinking, could still be sufficient
     - i.e. "if X passes the test then X thinks" might be true even if "if X thinks then X must be able to pass the test" may not

Criticisms of Turing's Position

2. The sense organs objection:
   - Turing Test only focusses on computer's ability to use words – no test of computer's ability to relate words to things in the world.
   - A computer that passes the TT but cannot supply the word "cup" when presented with a cup cannot be said to understand the words it uses.
   - Therefore test too weak and should be strengthened by insisting computer be given artificial sense organs.
   - However
     - can probe a speaker's understanding of many words without investigating their sensory interaction with environment
     - Nothing precludes computers with sense organs entering the competition
     - Turing noted that giving a computer sense organs and then subjecting it to ``an appropriate course of education" might be the best of way to prepare it for the test

Criticisms of Turing's Position

3. The simulation objection:
   - "A simulated X is not an X", e.g. simulated diamonds are not diamonds
   - A computer that passes the Turing test has shown that is a good simulation of a thinking thing – but that is not the same as being a thinking thing
     - In the imitation game if a man convinces the interrogator he is a woman, i.e. simulates a woman, doesn't follow that he is a woman

Criticisms of Turing's Position

3. The simulation objection (cont):
   - But note
     - Some simulated X's are X's (e.g. a simulated voice is still a voice)
     - Need to distinguish 2 types of simulation
       - Simulation_1 lacks essential features of what is simulated, e.g. simulated death, leather
       - Simulation_2 is just like what is simulated but is produced in a different way, e.g. artificial coal, simulated voice
   - Objection only holds if computer simulation of thinking is always a simulation_1 and never good enough to be a simulation_2
   - But question is whether a computer simulation of thinking could be a simulation_2 – simulation objection prejudges this

Criticisms of Turing's Position

4. The black box objection:
   - Turing test treats computer as a black box and looks only at output. A weakness?
   - Turing thought this was fine – treat our fellow human beings in the same way
   - Copeland thinks not: our judgement that our fellows think based not only on observation of their behaviour (outputs) but on fact of our biological similarity

Criticisms of Turing's Position

4. The black box objection (cont):
   - Suggests modifying Turing Test to insist
     - it pass the interrogation test (outward criterion)
     - it pass a test whereby we examine the program to see how output is produced and be convinced it is not simply a gigantic lookup table/ pattern matching chatbot (design criterion)
   - To meet design criterion program would have to either
     1. Work like a human (too anthropomorphic?)
     2. Be modular/extendible so could, e.g. be added to a robot with motor and sensory systems

Criticisms of Turing's Position

- Copeland goes on to suggest maybe Turing Test approach should be abandoned
  - Simply don't know how to assess the application of concepts like "thinking" to radically new sorts of entities like programmable computers
  - Suggests key features of creatures we are happy to apply the term "thinking" to are those whose action-directing inner processes are massively adaptable
    - E.g. can form plans, analyse situations, deliberate, reason, exploit analogies, revise beliefs in the light of experience, etc. in the real world
    - For such creatures it is reasonable to explain/predict their behaviour in terms of what the agent thinks/believes/wants, i.e. using what philosophers call intentional language
    - Thus, if we can build robots to which we are happy to apply the framework of intentional explanation/prediction, why not say they can think?

- Daniel Dennett on the Turing Test

Lecture Outline

- How can we decide if a machine is intelligent?
  - Turing Test
  - Arguments Against the Turing Test
- Is AI Possible?
  - Searle's Chinese Room Argument
  - Arguments Against Searle

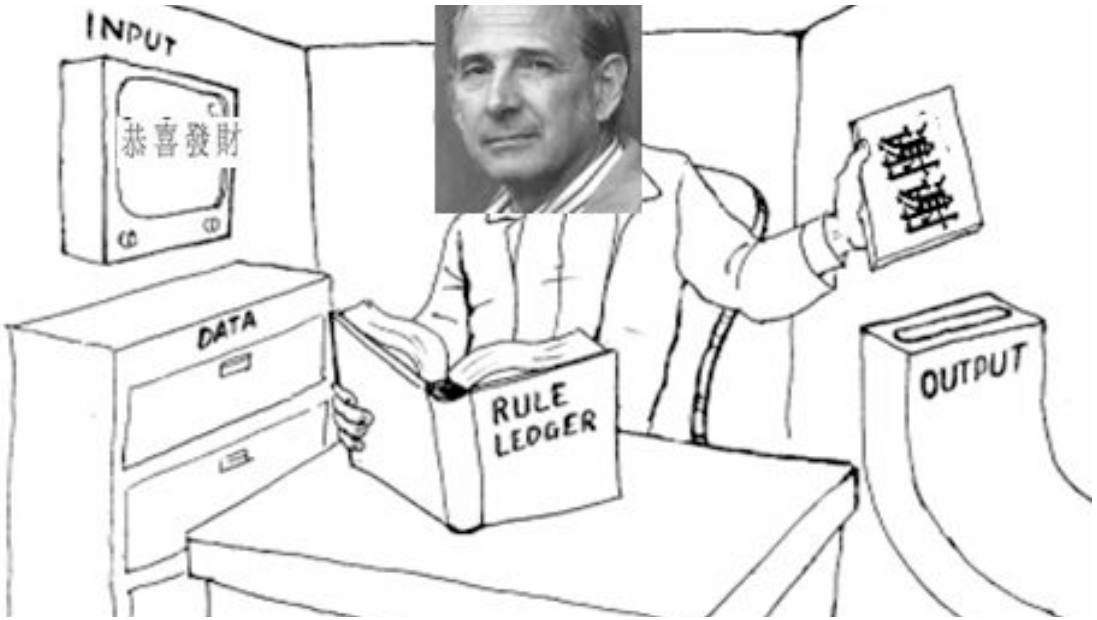
Is AI Possible?

- Depends on what we mean by AI …
- The question of whether the Strong AI hypothesis is true has led to extensive debate in the philosophical community.
- Central to this debate has been John Searle's Chinese Room thought experiment which he claims demonstrates the Strong AI hypothesis cannot be true …

Searle's Chinese Room Argument

- Searle asks us to imagine:
  - A man inside a closed room with two slots in the walls – one marked input, the other output
  - The man is seated at a table with a large rulebook and lots of blank paper and pencils
  - Through the input slot someone pushes questions written on paper in Chinese (we assume the man inside does not know Chinese)
  - Following instructions in the rule book the man
    - looks up the characters on the input slip
    - performs operations on them (e.g. translating them to binary strings)
    - finally arrives at a new Chinese string which he writes on a slip of paper and pushes through the output slot

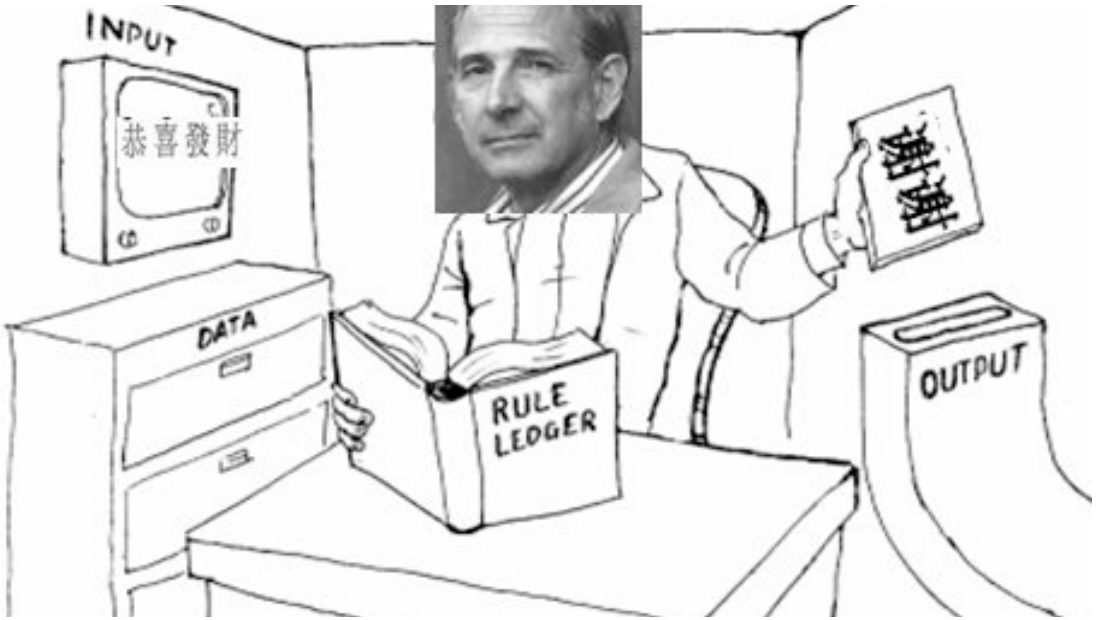
Searle's Chinese Room Argument



From: http://www.alexandria.nu/ai/blog/attachments/00000050_ChineseRoom.jpg

- Note: This scenario is NOT about translating between Chinese and some other human language (i.e. machine translation)
  - It's about taking action based on processing a string of symbols in a language one does not understand

Searle's Chinese Room Argument



From: http://www.alexandria.nu/ai/blog/attachments/00000050_ChineseRoom.jpg

- Suppose the Chinese room performs so well it passes the Turing Test
- Clearly the man inside does not understand Chinese
- But a computer answering questions does nothing more than the man in the room would do
- Therefore, a computer passing the Turing Test cannot be said to be thinking/understanding and thus the Strong AI hypothesis is false

Searle's Chinese Room Argument

- Note Searle claims the question (whether computers can think) is NOT an empirical question
  - I.e. is not one that can be settled by experimentation
    - So no future Turing Test has any relevance for the question "Can machines think?"
  - He claims his Chinese room argument settles the argument
    - Essentially no amount of syntactic manipulation can lead to semantics
- Is the argument valid?
- Many people have argued that it is not
  - See, e.g. Copeland (1993)

Copeland's Rebuttal

- Copeland begins by noting it's wrong to infer that because

  A: the man inside the room (call him Joe) does not understand Chinese

  therefore

  B: the Chinese room (slots, rule books, paper, man) does not understand Chinese)

- I.e. the system as a whole might understand Chinese even while Joe – the man inside – does not

Copeland's Rebuttal (cont)

- Searle replies to this point by asking who can believe the whole system understands Chinese
- Copeland's response is to observe that he was not claiming it could, but only that Searle's argument (because A therefore B) was invalid, not that B was true
- Copeland doesn't think B is true
  - but this is because the Chinese room is not a plausible system for passing the Turing Test, not because we can establish *a priori* that computers cannot think
- Argument explored in more detail in Copeland (1993) Chp 6

Summary

- First and most famous test for whether computers can think is the Turing Test
  - Based on the Imitation Game – game relying on conversation to identify gender of unseen participants
- Turing anticipated and refuted many potential objections to his proposed test
  - Copeland summarises other objections and discusses how the TT could be adapted to take them into account
- Searle's Chinese Room is an attempt to show by analysis alone that thinking machines are an impossibility
- Searle's arguments are rejected/ignored by the AI community, who continue to push at the limits of what capabilities we can bestow upon computers

Mandatory/Recommended Reading

- Turing, A.M. (1950). *Computing machinery and intelligence*. Mind, 59, 433-460.
- Wikipedia: Eugene Goostman. http://en.wikipedia.org/wiki/Eugene_Goostman
- Searle, John. R. (1980) *Minds, brains, and programs*. Behavioral and Brain Sciences 3 (3): 417-457.
- Russell, Stuart and Norvig, Peter (2010) Artificial Intelligence: A Modern Introduction (3rd ed). Pearson. Chapter 26.1-26.2.

References

Boden, Margaret (1977) Artificial Intelligence and Natural Man. Basic Books.

Copeland, Jack (1993) Artificial Intelligence: A Philosophical Introduction. Blackwells.

McCorduck, Pamela (2004) Machines Who Think. A K Peters/CRC Press.

Rich, Elaine and Knight, Kevin (1991) Artificial Intelligence (2nd ed). McGraw Hill.

Russell, Stuart and Norvig, Peter (2010) Artificial Intelligence: A Modern Introduction (3rd ed). Pearson.

Searle, John. R. (1980) Minds, brains, and programs. Behavioral and Brain Sciences 3 (3): 417-457.

Searle, John (1999) Mind, language and society, New York, NY: Basic Books.

Turing, Alan M. (1950) Computing Machinery and Intelligence. *Mind LIX(236)* , 433-460.

Whitby, Blay (2008): Artificial Intelligence: A Beginner's Guide. Oneworld Publications.

Wikipedia: Artificial Intelligence. http://en.wikipedia.org/wiki/Artificial_intelligence (visited 26/09/15).

Wikipedia: Turing Test. http://en.wikipedia.org/wiki/Turing_test (visited 29/09/15).

Wikipedia: Eugene Goostman. http://en.wikipedia.org/wiki/Eugene_Goostman (visited 26/09/15)

Wikipedia: Chinese Room. http://en.wikipedia.org/wiki/Chinese_room (visited 26/09/15).