Can Machines Think?

- How do minds work?
- Can minds work without bodies?
- Is it possible for machines to act intelligently like humans?
- If it is, can we say they have a mind?
- What criteria can we use whether an entity is intelligent or has a mind?

Can he think?

?

Can he think?

- performs a single task that is commonly said to require intelligence
- behaves like a human (to some extent)
  - in the domain of chess, people attribute human-like behavior to it ("*... What is he planning?...*")
  - even though it plays very different than a human player (e.g., it would never make certain typical mistakes)

Can he think?

- behaves like a human
  - it interacts, shows emotions, is capable of planning, can solve difficult problems, ...
- communicates like a human
  - not quite (at least not in a comprehensible language)

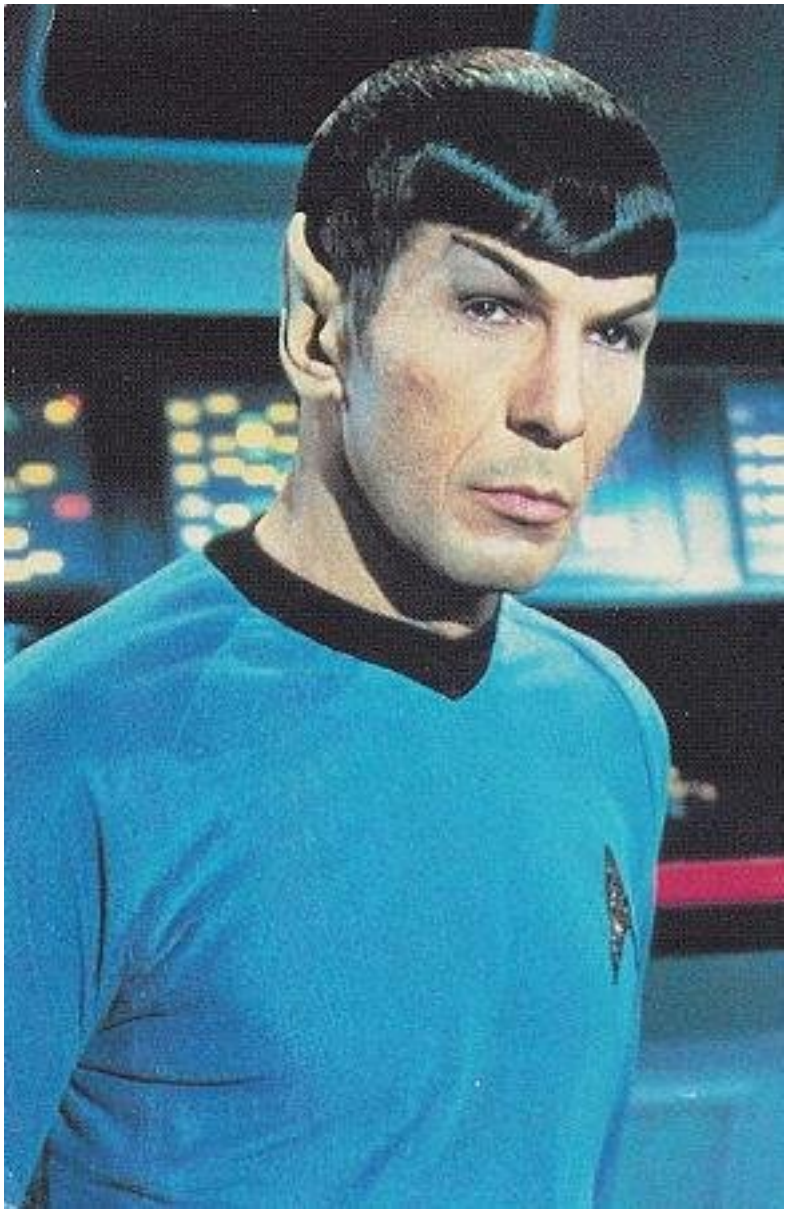
Can he think?

- communicates like a human
- behaves like a human
- looks like a human
  - not quite

Can they think?

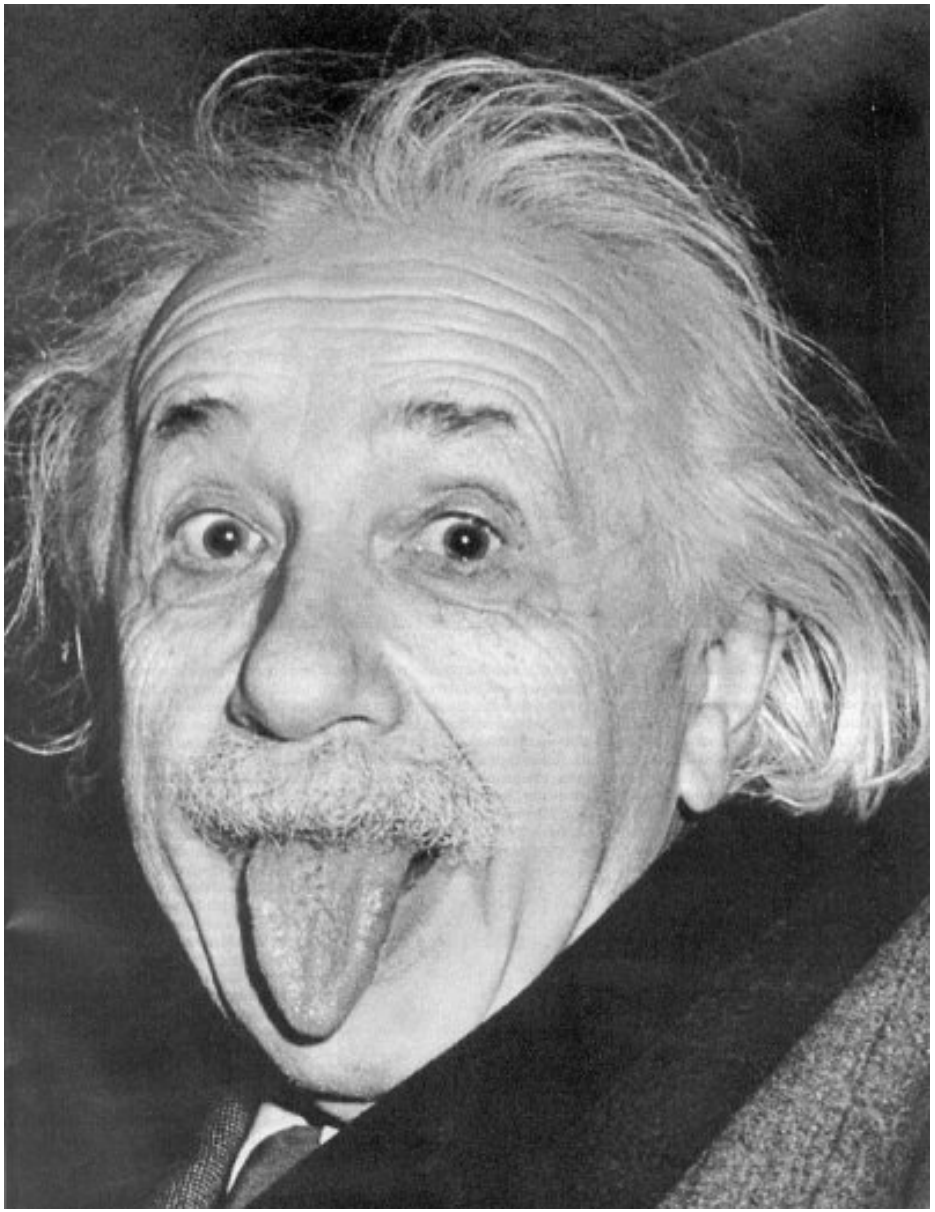
- communicate like a human
- behave like a human
- look like a human

Can he think?



- behaves like a human
- communicates like a human
- looks like a human
  (almost)
- has a brain like a human
  (presumably)

Can he think?



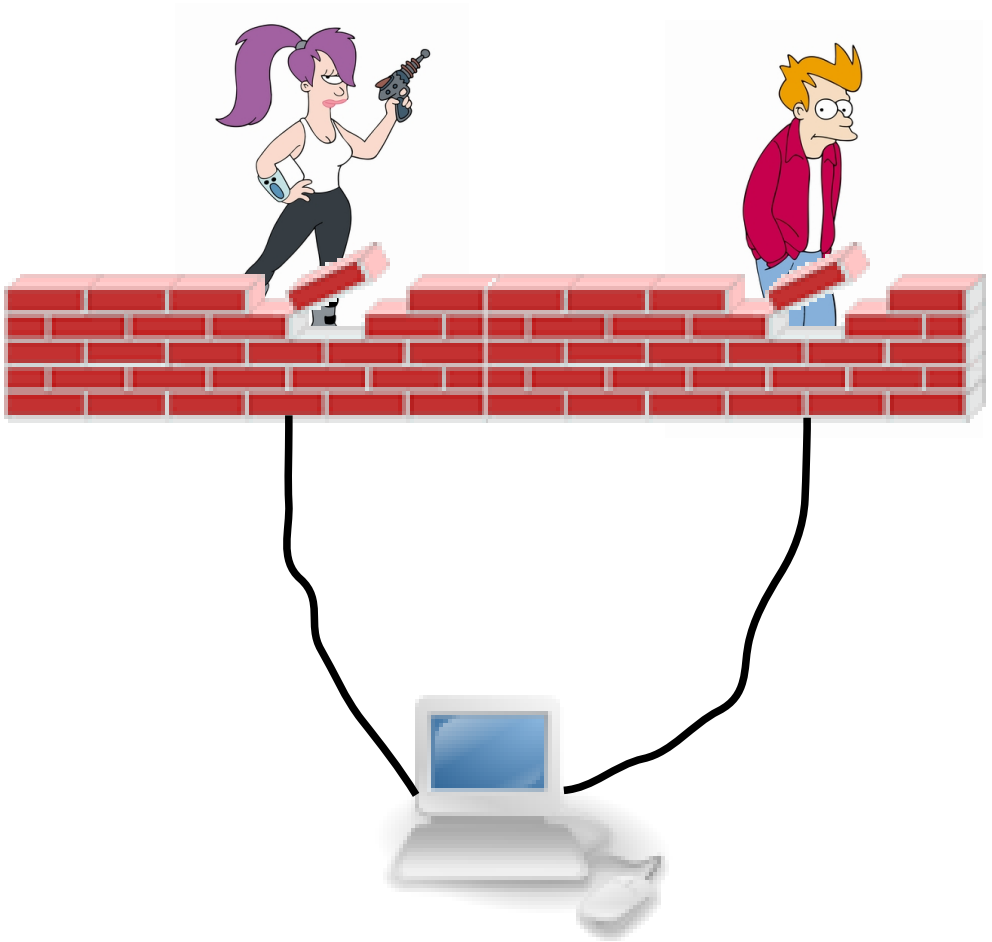
- behaves like a human
- communicates like a human
- looks like a human
- has a human brain
  (currently at McMaster University, Ontario, Canada)

Can he think?


I apologize for
this latest entry
I can't find a
chimp making
a face as dumb
as this one
-Bush

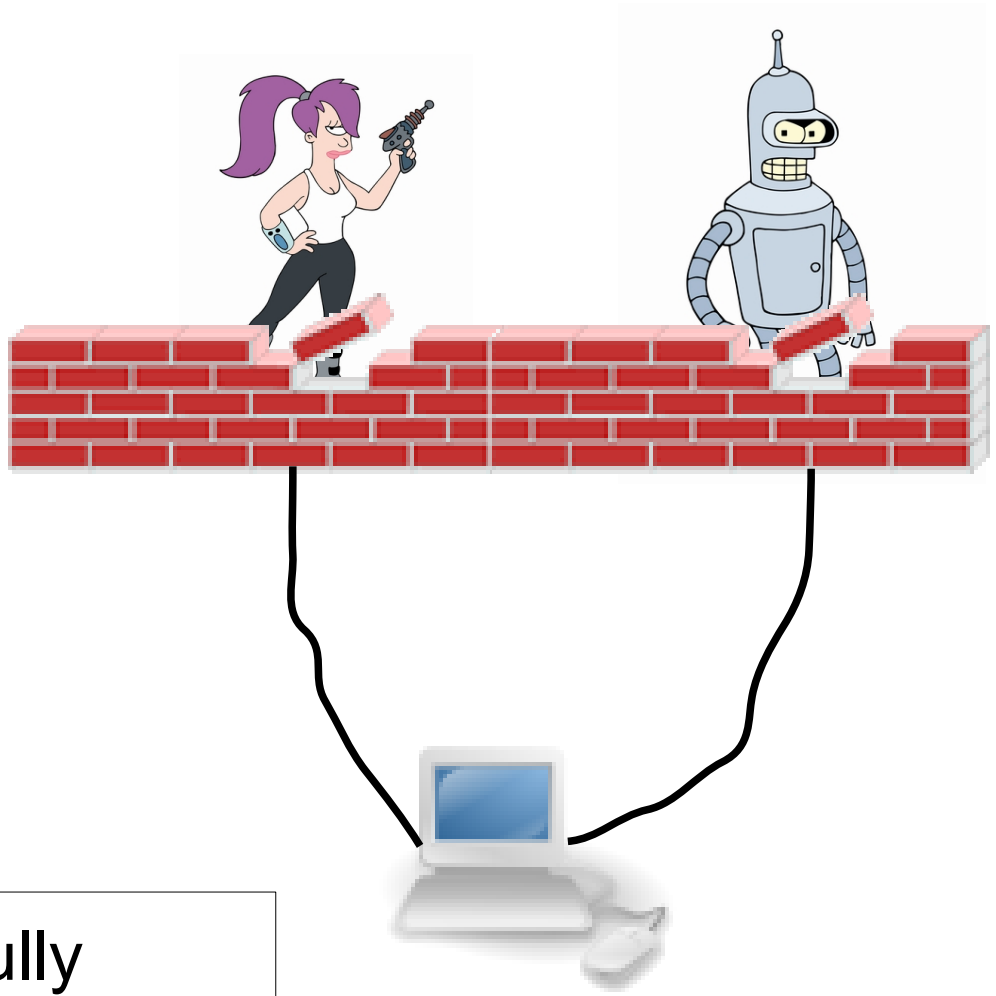
Imitation Game

- A human judge is connected (via a teletyper/computer) to a man and a woman
  - the judge can ask any question about any subject
  - the woman replies honestly
  - the man tries to pretend he is a woman
- The judge has to find out:
  - Who is the woman?



Turing Test

- A human judge is connected (via a teletyper/computer) to a computer and a woman
  - the judge can ask any question about any subject
  - the woman replies honestly
  - the computer tries to pretend it is a woman
- The judge has to find out:
  - Who is the woman?



- If the computer can successfully fool the judge, it has passed the test
  - it can be considered intelligent

Turing Test

- Basic Idea:
  - Instead of defining a long and controversial list of necessary prerequisites for intelligence
  - compare to undeniably intelligent beings → humans

"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 5 minutes of questioning"

-Alan Turing (1950)

Alan Mathison Turing (1912-1954)

Hypothetical TT Transcript

Q: Please write me a sonnet on the subject of the Forth Bridge.

A: *Count me out on this one. I never could write poetry.*

Q: Add 34957 to 70764

A: (Pause about 30 seconds and then give as answer) *105621.*

Q: Do you play chess?

A: *Yes.*

Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?

A: (After a pause of 15 seconds) *R-R8 mate.*

Eliza

- Program by Joseph Weizenbaum (1966)
  - was meant to be a parody on a Rogerian psychoanalyst
  - takes excerpts from the subject's comments and poses questions back to the subject
  - works by very simple pattern matching on responses and canned phrases
  - has absolutely no understanding of the conversation
- Some people claim that ELIZA was the first program to pass the Turing Test because it often fooled people
  - Weizenbaum's intention was not to demonstrate AI
  - he later devloped into a strong critic of computer technology
- Many implementations
  - Emacs: `M-x doctor`
  - AOLiza http://fury.com/aoliza/

The Loebner Competition

- Modern day version of the Turing Test
  - http://www.loebner.net/Prizef/loebner-prize.html
  - programs and humans converse about a specific topic of the program's choice
  - multiple judges rank-order multiple humans and multiple computer programs from ‘most likely to be human’ to ‘least likely to be human’.
- Loebner has promised $100,000 for the first computer program to be ‘indistinguishable from a human’.
  - minor prizes for best program, etc.

- Interesting competition, but does not advance state-of-the-art in AI
  - most programs apply cheap tricks like ELIZA
  - test scenario is too restrictive to be conclusive (one topic only)

CAPTCHAs

- Completely Automated Public Turing test to tell Computers and Humans Apart
- Turing test backwards
  - because we know that computers cannot pass a Turing test, we give them simple questions that can only be answered by humans
    - purpose: make sure that Web bots do not register a million free E-mail accounts, etc.

- Examples:
  - What is written here?
    (Computers can't see)

  - Please calculate 53 minus 11
    (Computers can do the math, but don't understand the question)

Strong vs. Weak AI

- Weak AI
  - build agents that act rationally
  - accomplish specific problem solving or reasoning tasks
    - e.g., playing chess, mowing the lawn, solve a Sudoku puzzle, ...
  - but are not universally intelligent
    - Can they think?
- Strong AI
  - Build a universally intelligent agent
    - encompasses the full range of human cognitive abilities
    - can pass the Turing test
  - AI-complete problems:
    - informal (and half-joking) term denoting problems for which it is believed that universal intelligence is necessary
  - Philosophical discussion: Is Strong AI possible?

Physical Symbol Systems Hypothesis

- Physical Symbol Systems
  - ”A ***physical symbol system*** consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). A symbol system also possesses a number of simple processes that create, modify, copy and destroy symbols. [...]”
- The Physical Symbol System Hypothesis
  - ”A physical symbol system has the necessary and sufficient means for intelligent action.” *(Newell & Simon, 1970)*
- PSS and Strong AI
  - PSSH is at the heart at what is known as "strong AI"
  - annoys many philosophers (such as Searle) and many humanists who find the concept degrading to their sense of what it is to be human.

9 Objections to Strong AI

Turing himself started the philosophical debate by formulating 9 possible objections to intelligent machines:

- Theological objection:
  - Thinking is part of humans' souls, and so animals/machines can't think.
  - Turing's answer:
    - theological arguments do not impress him in general
    - If God wishes, he could give a soul to a machine, couldn't he?
- Head-in-the-sand objection:
  - Consequences of thinking machines are terrible, so let's hope it's not possible.
  - Turing's answer:
    - not substantial enough to require refutation

9 Objections to Strong AI

- Mathematical Objection
  - Gödel's Incompleteness Theorem:
    - in any consistent logical system that includes number theory, there are statements that can't be proved or disproved
    - The trick: one can formulate a sentence like
      "This sentence cannot be proved."
  - Lucas (and Penrose):
    - Machines are formal systems, so there will be a formula the machine will be unable to produce as true, although a mind can see that it is true. And so the machine will not be an adequate model of the mind.
  - Answers:
    - Humans are also fallible
    - The Gödel trick can also be applied to humans:
      "J. R. Lucas cannot assert that this sentence is true"

9 Objections to Strong AI

- Consciousness Objection
  - *"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."*
  - Turing's answer:
    - We cannot be certain that other people think or feel, it is a polite convention to assume that everyone thinks and feels.
    - Why not extend this courtesy to machines if they show evidence of their feelings?
      → following possible TT dialogue

Interrogator: In the first line of your sonnet which reads *'Shall I compare thee to a summer's day'*, would not *'a spring day'* do as well or better?

Witness: It wouldn't scan.

Interrogator: How about *'a winter's day'* That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

9 Objections to Strong AI

- Objections from various disabilities:
  - I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do X
    - X = {*Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make some one fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really newbe self-aware, have sense of humor, fall in love, etc.*}
  - Turing's answer:
    - He thinks that these need to be investigated first
    - but does not see any particular reason why they could not be done.

9 Objections to Strong AI

- Lady Lovelace's objection
  - Lady Lovelace (Ada Byron) wrote about Babbage's Analytical Engine:
    - "It has no pretensions to *originate* anything. It can do *whatever we know* how *to order it* to perform."
  - In other words:
    - Computers can only do what they are programmed to do and cannot surprise us
  - Turing's answer:
    - What about learning machines?
    - Humans often are surprised by machines.
  - Similar fallacy:
    - W. Ross Ashby: Can a Mechanical Chess-Player Outplay Its Designer? *The British Journal for the Philosophy of Science 3*(9):44-57, 1952.
    - His answer to the question in the title:
      - No, because it can only replicate the thoughts of its programmer

9 Objections to Strong AI

- Continuity of the Nervous System
  - The nervous system is not a discrete-state machine, so it can't be modeled by a computer
  - Turing's answer:
    - It can be approximated well enough
  - this is still heavily debated
    - e.g.: Physicist Roger Penrose believes that consciousness is due to *quantum gravity*

9 Objections to Strong AI

- Argument from Informality of Behaviour
  - *"If each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines"*
  - Turing's answer:
    - the determinism lies deeper
      - at least we follow laws of physics
      - we also cannot be sure that there are no higher laws of behavior that we follow
- Argument from Extra-Sensory Perception
  - Turing suggest a telepathy-proof room...
    (otherwise the judge could distinguish because of ESP)

The AI debate

- After the birth of AI (1956), a heated discussion about potential success and limitations of AI emerged
  - Herbert Simon, 1958:
    - "within ten years a digital computer will be the world's chess champion."
  - Hubert Dreyfus, 1972: What Computers Can't Do
    - Human intelligence is more than manipulation of symbols.
    - very aggressive and highly controversial book
  - John Searle, 1980:
    - "Chinese Room" thought experiment
    - Opposed idea of strong AI, that machines can think

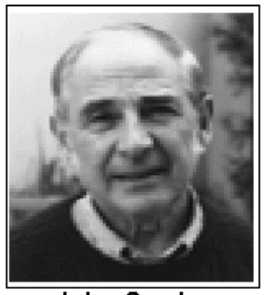
Searle's Chinese Room

Classical thought experiment by John Searle (1980) with the goal of defeating strong AI

- Scenario from the outside:
  - there is a room with which you can lead an intelligent conversation in Chinese
  - by submitting a question (or any other statement) on paper through a slot in the door
  - and receiving an answer through the same slot
- Scenario from the inside:
  - The man in the room does not speak or understand Chinese
  - He has a set of rules (in English) that he has to follow
  - Upon receipt of a query, he applies his English rules to manipulate the symbols on it and eventually produces a response
  - The response is returned through the slot in the door

Searle's Chinese Room

3 John Searle, 1980a, 1980b, 1990b
The Chinese Room argument. Imagine that a man who does not speak Chinese sits in a room and is passed Chinese symbols through a slot in the door. To him, the symbols are just so many squiggles and squoggles. But he reads an English-language rule book that tells him how to manipulate the symbols and which ones to send back out. To the Chinese speakers outside, whoever (or whatever) is in the room is carrying on an intelligent conversation. But the man in the Chinese Room does not understand Chinese; he is merely manipulating symbols according to a rule book. He is instantiating a formal program, which passes the Turing test for intelligence, but nevertheless he does not understand Chinese. This shows that instantiation of a formal program is not enough to produce semantic understanding or intentionality.
Note: For more on Turing tests, see Map 2. For more on formal programs and instantiation, see the "Is the brain a computer?" arguments on Map 1, the "Can functional states generate consciousness?" arguments on Map 6, and sidebar, "Formal Systems: An Overview," on Map 7.


John Searle

in · ten · tion · al · it · y: The property (in reference to a mental state) of being directed at a state of affairs in the world. For example, the belief that Sally is in front of me is directed at a person, Sally, in the world. Intentionality is sometimes taken to be synonymous with representation, understanding, consciousness, meaning, and semantics. Although there are important and subtle distinctions in the definitions of "intentionality," "understanding," "semantics," and "meaning," in this debate they are sometimes used synonymously.

The Chinese Room Argument

- The Chinese Room would pass a Turing Test
  - from the outside, you have to assume that there is a fluent Chinese speaker hidden in the room
- However, nothing in the room knows Chinese
  - the symbol manipulator just follows rules that are written in plain English
- Therefore:
  - Just as the room does not know any Chinese, a computer that passes the Turing Test can not really think
  - The Turing Test cannot be used to discover “true” intelligence, it can only be used to discover “simulated” intelligence
- Intentionality:
  - Searle's word for the difference between “true” intelligence and “simulated” intelligence

Purpose of the room

- Show that the Turing test is inadequate.
- Debunk the claims of strong AI
  - Searle claims that because no part of the Chinese room has understanding (i.e. doesn't know what the input means) it is clear that just relying on syntax to preserve semantics doesn't explain what it is to be a mind
- Understanding as simulation:
  - We don't think something is really burning when we simulate a fire. Why think we are dealing with a mind when we simulate intelligence?
- Nevertheless, Searle thinks that machines can have understanding
  - Humans are an example
  - but formal descriptions of machines are neither necessary nor sufficient for intelligence

Replies to Searle

- Systems reply
  - The human in the Chinese room is only one part of the system – it is the entire system (including the room itself) that understands Chinese
  - Searle's Response:
    - The human can memorize the rules, and then he would appear to communicate in Chinese, although he still operates in English
- Robot reply
  - Add a camera, and manipulators, and relate the formal symbols to objects in the real world (*symbol grounding*)
  - Searle's Response:
    - Still no intentional states – where would they be?
    - Robot response acknowledges need for more than formal symbol manipulation.

Replies to Searle

- Brain simulator reply
  - Simulate the sequence of neuron firings that occurs in the brain of a native Chinese speaker
  - Searle's Reply:
    - This concedes that Strong AI (PSSH) is not possible
    - You can simulate the neurons with waterpipes, neither the operator nor the waterpipes understand Chinese
- Combination reply
  - Put a simulated brain into a body
  - Searle's Reply:
    - Once we learned how the system functions we will know that it has no intentionality

Replies to Searle

- Other Minds Reply
  - Also for people, we can only assume that they have minds by observing their behavior
  - Searle's reply:
    - In "cognitive sciences" one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects.
- Many Mansions reply
  - Maybe there are other ways to achieve intentionality than via programming?
  - Searle's Reply:
    - This misses the point. Strong AI claims that it is possible by programming.

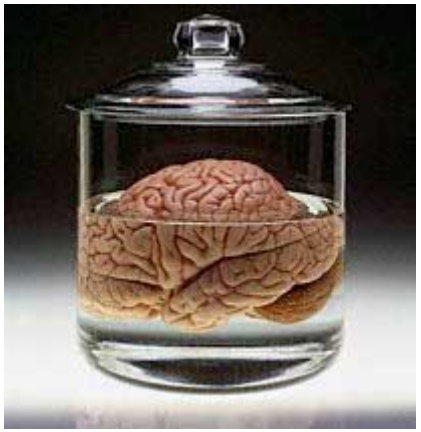
The Mind-Body Problem

- asks how mental states and processes are related to brain states and processes
  - very old philosophical (and religious) problem
- **Dualism:**
  - Mind and Matter are two different things
    - Mental phenomena are non-physical
  - Clearly formulated by Descartes (17th cent.)
    - your thinking self does not have an extension
    - mind and body influence each other
  - Mind = Self, Personality, Soul
    - consciousness, self-awareness

- **Monism** (or Materialism):
  - Mental states are brain states
  - Searle: *"Brains cause Minds"* (i.e., we need the right hardware)
- Strong AI question: Can minds be built on computer hardware?

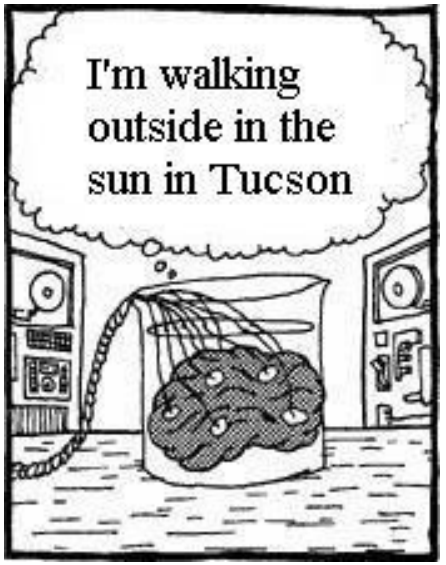
The Brain-in-a-Vat Argument

- your brain is removed and put in a vat
  - all neural connections are replaced with bluetooth transmitters



- will it still work?
  - will you still be conscious?
  - can you stand in front of your brain and watch yourself think?

→ The Matrix



The Brain-Prosthesis Experiment

- Scenario:
  - Assume that neurons can be replaced with electrical circuitry
  - replace, neuron by neuron, the entire brain
  - then revert the process, putting your neurons back in place
  - so after the operation you are (physically) the same as before
- Purpose:
  - Nobody can tell from the outside whether somebody is self-conscious or not
  - But in this case you should be able to report this experience
- Searle holds a red object before you and asks what you see:

> *You find, to your total amazement, that you are indeed losing control of your external behavior. ... You want to cry out "I can't see anything. I'm going totally blind." But you hear your voice saying in a way that is completely out of your control, "I see a red object in front of me."*

- The other side (Moravec) think they'd be perfectly conscious

So do computers think?

- In many domains they at least make the impression.
- Garry Kasparov, 1996, after 1st Deep Blue Match (he won)
  - *I got my first glimpse of Artificial Intelligence on Feb. 10, 1996, at 4:45 p.m. EST, when in the first game of my match with Deep Blue, the computer ... [made] a wonderful and extremely human move.*
  - *I could feel - I could smell - a new kind of intelligence across the table. While I played through the rest of the game as best I could, I was lost; it played beautiful, flawless chess the rest of the way and won easily.*
  - *If the computer makes the same move that I would make for completely different reasons, has it made an "intelligent" move? Is the intelligence of an action dependent on who (or what) takes it?*
  - *So although I think I did see some signs of intelligence, it's a weird kind, an inefficient, inflexible kind that makes me think I have a few years left.*

Garry Kasparov, The Day That I Sensed a New Kind of Intelligence, *Time Magazine* 147, 13 (March 25 1996).
http://www.time.com/time/magazine/article/0,9171,984305-1,00.html

So do computers think?

- But one year later:
- Garry Kasparov, 1997, after 2nd Deep Blue Match (he lost)
  - *"It was nothing to do about science, it was one zeal to beat Garry Kasparov. And when a big corporation with unlimited resources would like to do so, there are many ways to achieve the result."*
  - *"I feel confident that the machine's win hasn't proved anything. It's not yet ready, in my opinion, to win a big contest."*

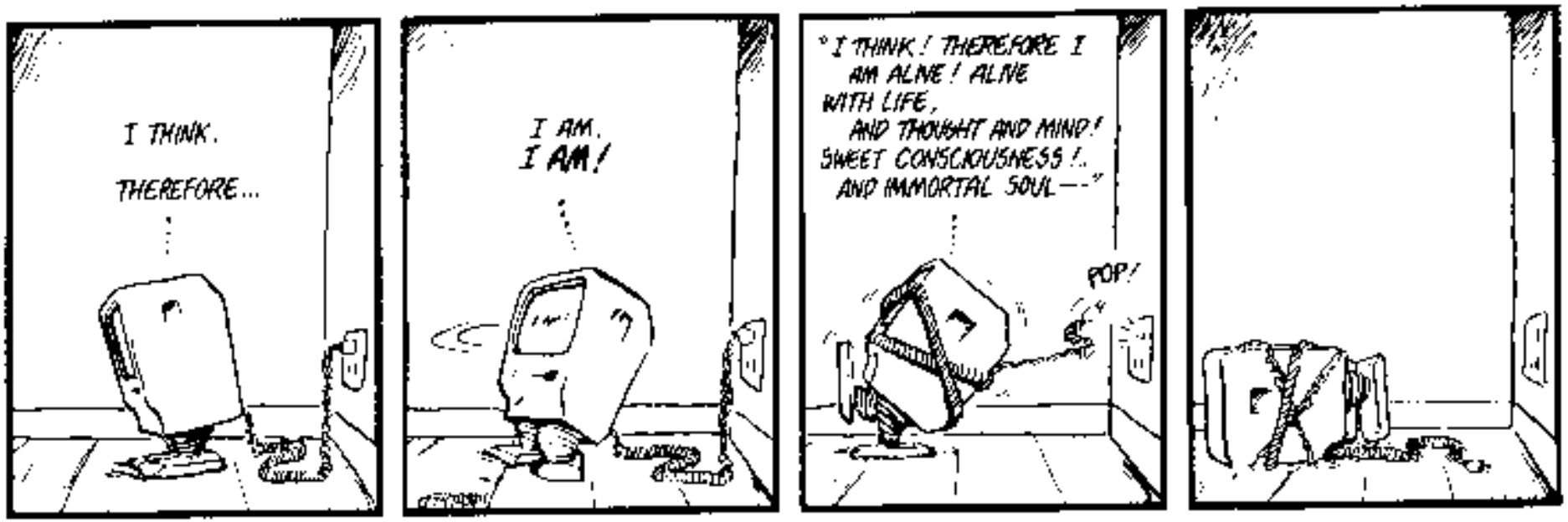
http://www.cnn.com/WORLD/9705/11/chess.update/

- Note:
  - Kasparov lost game two (and later the match) because he was completely baffled by a human-like move, lost his nerves, and eventually accused the IBM team of cheating.

In Turing's Words

"The original question, 'Can machines think?', I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."

-Alan Turing (1950)


I THINK.
THEREFORE...
I AM.
I AM!
"I THINK! THEREFORE I AM ALIVE! ALIVE WITH LIFE, AND THOUGHT AND MIND! SWEET CONSCIOUSNESS!.. AND IMMORTAL SOUL—"
POP!

Recommended Reading

- Alan M. Turing (1950), “Computing Machinery and Intelligence”, *Mind* 59:433-460.
  - http://www.loebner.net/Prizef/TuringArticle.html
- Allen Newell and Herbert A. Simon (1976), “Computer Science as Empirical Inquiry: Symbols and Search”, *Communications of the ACM 19*(3). (ACM Turing Award Lecture 1975)
  - http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/PSS/PSSH1.html
- John Searle (1980), “Minds, Brains, and Programs”, *Behavioral and Brain Sciences* 3:417-424.
  - http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html
- Joseph Weizenbaum (1966). “ELIZA - a computer program for the study of natural language communication between man and machine”, *Communications of the ACM* 9(1): 36-45.
  - http://i5.nyu.edu/~mm64/x52.9265/january1966.html