Lecture 23: Conclusions. Philosophy of AI

- What we covered and left out in the class
- Can computers become intelligent?
  - Turing test revisited
  - Strong AI vs weak AI
- The singularity (& beyond?)

Major topics we talked about

- Search: crucial! Many other AI algorithms rely heavily on search or can be viewed as search
- Logic as a form of knowledge representation that supports inference
  Unfortunately, the world is usually not logical
- Planning: how to decide what to do to achieve a goal?
- Probabilities and Bayes nets: A more flexible form of knowledge representation, still allows inference, can be learned form data
  But may be less intuitive for people
- Utilities, decision making under uncertainty and for sequential problems
  Strong connections to psychology, economics
- Machine learning: how do we learn new things from data?
  Crucial for intelligence!

Some key concepts

- Heuristics
- Local vs global methods
- Efficient data structures are really important!
- Error functions and gradients
- Cross-validation

We did not talk about: Robotics

- Motion planning (very similar to search)
- Sensor processing (laser, IR, sonars, cameras...)
- Lots of tasks are being tackled!
  - Nursing/helping disabled people
  - Driving
  - Space exploration
  - Remote surgery
  - Entertainment
- See also COMP-417

We did not talk about: Vision/Perception

- Object recognition
- Song recognition
- Speech recognition
- Image segmentation: what objects are in an image?
- Face recognition
- Video analysis/activity recognition
- See also COMP-557...

We did not talk about: Multi-agent systems

- Cooperative situations: all agents work toward one common goal
  - Well understood, can get teams of agents programmed to cooperate
  - Can even make them learn to cooperate (e.g. by sharing reward)
- Self-interested agents: simulates economy-related situations
  - Auctions
  - Supply-chain management
  - Emergent behavior in agent societies (where each agent behaves according to very simple rules)
- Emerging: social networks
  - Use AI tools like machine learning, probabilistic models, natural language processing to understand how human social networks form, and how information propagates
  - Lots of data (Facebook, Twitter, blogs, ...) but interesting questions are not really related to computer science...

Why multi-agent systems are difficult: Prisonner's dilemma

- Two suspects are arrested by the police
- Police have insufficient evidence, so they offer a deal to each prisoner separately:
  - If A testifies against B, A walks free and B goes to jail for 5 years; similarly for B
  - If both remain silent, they both go to jail for 1 year
  - If each betrays the other, they both go to jail for 10 years
- Each prisoner must choose if they stay silent or betray the other (they cannot communicate)
- Each one is assured that the other would not know about the betrayal before the end of the investigation.
- *How should the prisoners act?*

Computers vs Minds

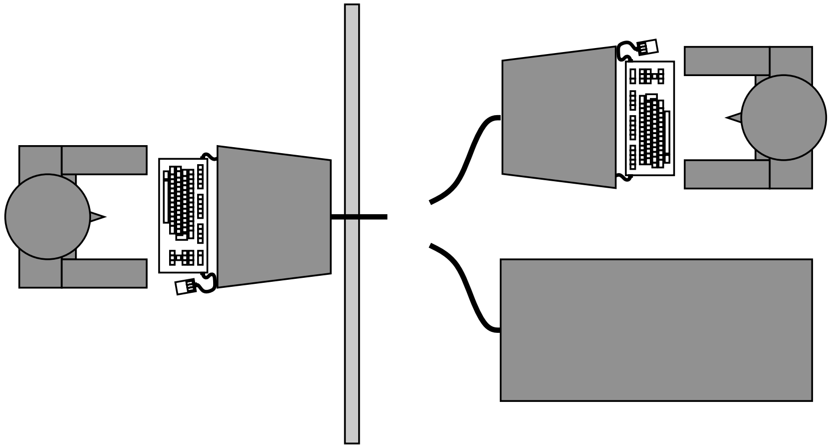
Computers are...	Biological beings are..
Dead	Alive
Programmed (for the most part)	Exhibit free will (for the most part)
Simple (in theory)	Complex (in theory)
Can process lots of data very fast	Slow but intelligent
Mostly deterministic	Creative

What can AI hope to achieve in a computer?

How to understand the difference between computers and minds?

- External factors:
  - “Intelligent”/purposeful behavior
  - Adaptation to changes
  - Interaction/appropriate response to the world (e.g. conversation)
  - Observation of “thinking processes” (e.g. fMRI)
- Internal factors:
  - Consicousness
  - Intentionality (I did something because I wanted to do it)
- The internal factors are considered crucial by some, but much harder to assess

Recall: Turing test



An operator interacts with either a human or an AI agent. Can he correctly guess which one?

Modern version: CAPTCHA

- Used to tell apart humans and bots on the internet
- Increasingly, character recognitions software is becoming able to handle these
- Much of the technology used to “break” CAPTCHAs is based on machine learning

Objections to computers passing the Turing test

- Theological: only immortal souls can think
- ‘Heads in the sand: dreadful consequences if we push computers to it!
- Mathematical: Godel’s Incompleteness Theorem

  But humans might be incomplete, too!
- Arguments from disabilities: Machines cant do X (humor, reflect, make mistakes, etc.)

  With enough memory, machines could be able to “fake” it
- Continuity of nervous system

  Real numbers can be approximated
- Informality of behavior: we dont follow rules

  Simple programs can seem unpredictable (by randomness, or by making decisions in a way that is counterintuitive to people)

Turing's conclusion regarding AI

"We may hope that machines will eventually compete with men in all purely intellectual fields. [...] Many people think that a very abstract activity, like playing chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. [...] Again I do not know what the right answer is, but I think both approaches should be tried. We can only see a short distance ahead, but we can see plenty there that needs to be done.

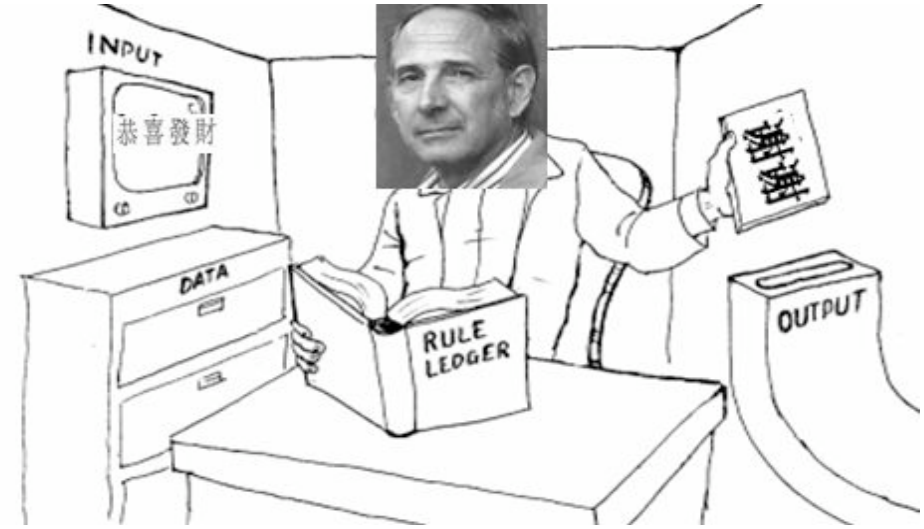
Weak vs. Strong AI

- *Weak AI*
  - We can build machines that act as if they were intelligent
  - We can study human intelligence by building computer models of it
  - Most AI research is in this area (and most researchers would agree that we are succeeding to a large extent)
- *Strong AI*
  - The goal is to build machines that are *actually thinking* “like people” (as opposed to just simulating thinking)
  - AI researchers do not care much about this, but philosophers do!
  - Some debate regarding whether strong AI is possible...

The Chinese room argument

- Cf. Searle (1980), “Minds, brains and programs”
- Suppose you have a person locked in a room.
- The person receives pieces of papers with Chinese symbols on it
- The person has a big rule book and a bunch of paper; by following the instructions in the book, the person can produce a new piece of paper with Chinese symbols on it
- The person does not know Chinese
- But from the *outside observer’s point of view*, the room appears to speak perfect Chinese!

The Chinese room argument



Conclusion: Might behave intelligently, but isn't!

- Argument used to show that a mere symbol processing machine cannot attain "understanding" or "intentionality" no matter how intelligent it appears to behave
- The Chinese room is not "conscious" or "intelligent" even though it appears so

Searle's later rebuttal of Chinese room

- Consciousness is an *emergent property* of a set of objects (neurons?)
- So maybe the Chinese room really understands Chinese, even though none of its parts do
- Just like "solidity" is an emergent property of a set of molecules and interactions...
- Now suppose that you took a brain and started replacing biological neurons with artificial ones: at what point does consciousness cease to exist?

Ethics for AI

- Robots are an increasing part of daily life, including for performing mundane tasks, or dangerous tasks

  Eg. Nursing care robots, "maid" robots, heavy-duty industrial automation, military drones

- "Robot ethics": what laws should robots obey?

  Cf. Asimov: protect humans, obey orders and protect themselves (in order)

- Might preclude us from using some technologies, which work very well but would not allow one to interpret the decisions that were made.

The singularity

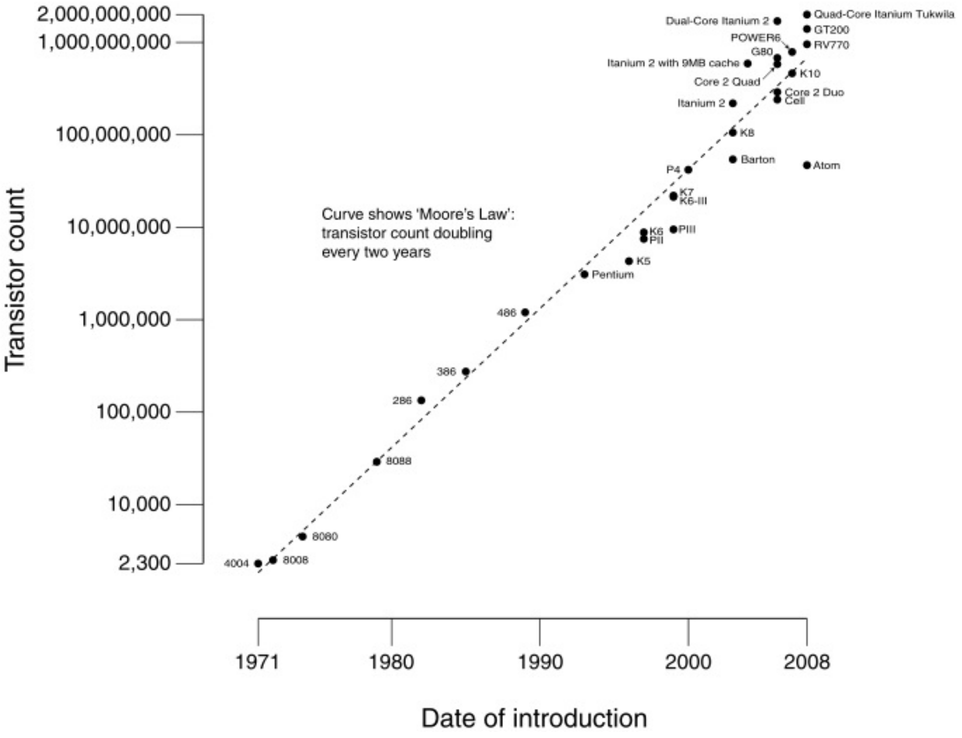
- Cf. John von Neumann: the ever-accelerating progress of technologygives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.
- Some people believe that at some point (during your life time!) computer technology will reach a point beyond which current models of prediction do not apply
- Most consider the "tipping point" to come from the creation of strong AI
- Possible consequences:
  - AIs will take over!
  - Humans will go into a "mind melt" with computers, and some kind of post-human super-race will emerge

    E.g. What if you could download all your brain (neurons and synaptic connections) onto a computer?

Can the singularity happen?

- Moore's law:

CPU Transistor Counts 1971-2008 & Moore's Law



- Similar exponential scaling has been observed in many other areas of technology (nr. of operations per sec per $1000, memory etc)

Is the singularity upon us now?

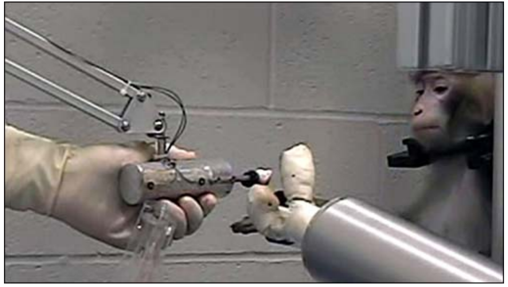
Cf. Hilbert & Lopez, Science, 2011

- "In 2007, humankind was able to store 2.9×10^{20} optimally compressed bytes, communicate almost 2×10^{21} bytes, and carry out 6.4×10^{18} instructions *per second on general-purpose computers*"
- General-purpose computing capacity grew at an annual rate of 58%.
- Telecommunication has been dominated by digital technologies since 1990 (99.9% in digital format in 2007)
- The majority of our technological memory has been in digital format since the early 2000s (94% digital in 2007).

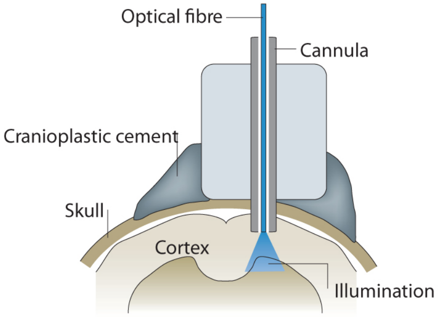
Is the singularity upon us now?

- Brain-computer interfaces: devices that communicate directly with a certain part of the brain

  E.g. monkey is moving a robotic arm by thought



- Optogenetics: controlling the brain with rays of light


Optical fibre
Cannula
Cranioplastic cement
Skull
Cortex
Illumination