Ethics

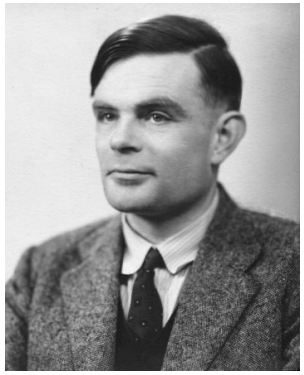
Larry Holder
School of EECS
Washington State University

Philosophical Issues

- Weak AI
  - Machines can act as if they were intelligent
- Strong AI
  - Machines can actually be intelligent (i.e., think)
- Can we tell the difference?
- Is even weak AI achievable?
- Should we care about achieving strong AI?
- Are there ethical implications?

Weak AI?

- Turing Test
  - Can the machine convince a human that it is human via written English
- Loebner Prize
  - en.wikipedia.org/wiki/Loebner_Prize
- AI XPRIZE (ai.xprize.org): $5M
- Mitsuku (mitsuku.com)

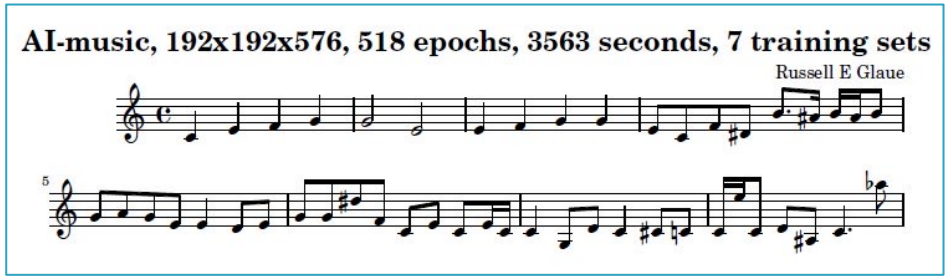


Alan Turing
(1912–1954)

The Singularity Is Near (2012)

Arguments Against Weak AI

- Disability
  - But a machine can never…
    - Beat a master at chess (✓)
    - Compose a symphony (~)
    - Laugh at a joke
    - Appreciate beauty
    - Fall in love



- Response
  - Magenta Project (magenta.tensorflow.org)
  - Engineer different approaches (planes vs. birds)
  - If we can understand how humans do it…

Arguments Against Weak AI

- Mathematical objection
  - Godel's incompleteness theorem
    - In any formal system there are true sentences that cannot be proven
    - "This sentence is not provable" is true, but not provable
- Response
  - Formal systems are infinite, machines are finite
  - Inability to prove obscure sentences not so bad
  - Humans have limitations too

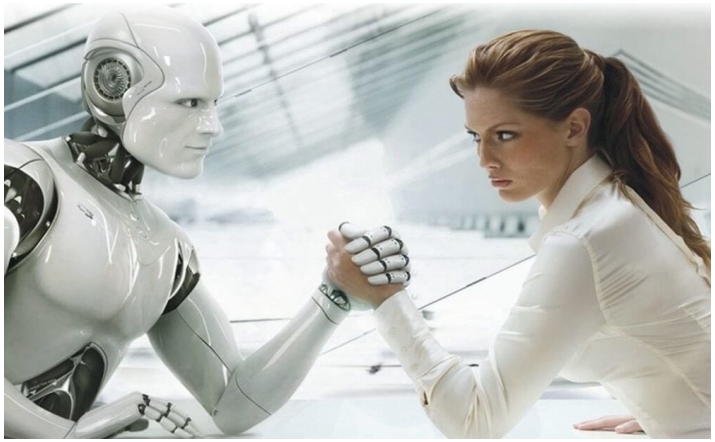
Kurt Godel
1906-1978

Arguments Against Weak AI

- Informality
  - Human behavior too complex to model formally
- Response
  - Usually assumes overly-simplistic models (e.g., propositional logic)
  - Learning can augment the model

Strong AI?

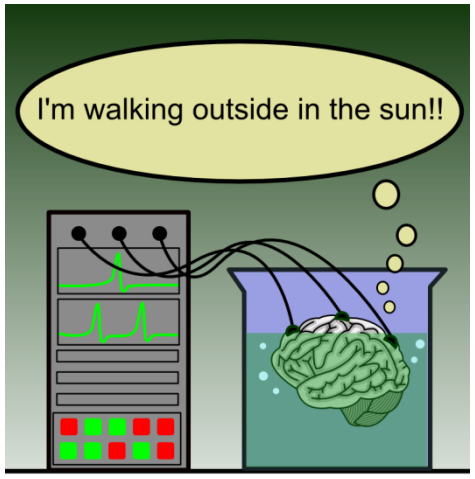
- Machine thinks like a human
- How do we define human thinking?
  - Machine has to know it passed the Turing test
  - Consciousness argument
- Mental state = physical (brain) state
- Mental state = physical state + ?
- Arguments ill-defined
- What is consciousness?



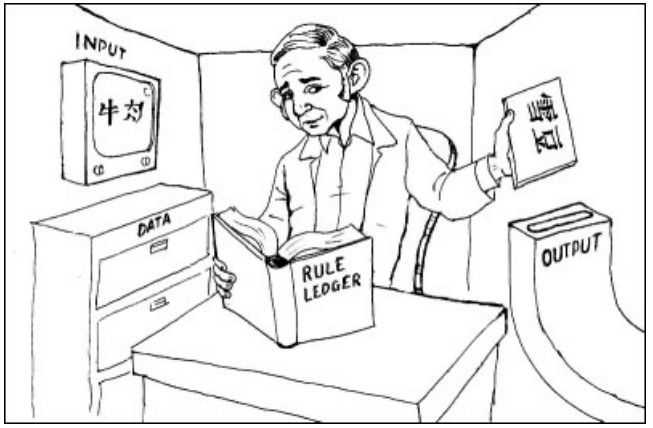
Strong AI?

- Functionalists say "Yes"
  - Brain maps inputs to outputs
  - Can be modeled as a giant lookup table
  - Brain in a vat
- Naturalists say "No"
  - Lookup tables are not intelligent
  - Searle's Chinese room argument
- Does achieving strong AI matter?

Brain in a Vat



Searle's Chinese Room



Ethical Issues

- Impact on economy: Losing jobs to automation
- Lethal and autonomous robots
- Surveillance and privacy
- Data mining

"Eagle Eye" (2008)

"Person of Interest" (2011–2016)

Ethical Issues

- AI responsibility
  - Generally, human experts are responsible for relying on AI decisions
  - Autonomous AI liability falls to the human designers
  - Can an AI system be charged with a crime?

"I, Robot" (2004)

Ethical Issues

- Stephen Hawking (2014)
  - “Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last.”

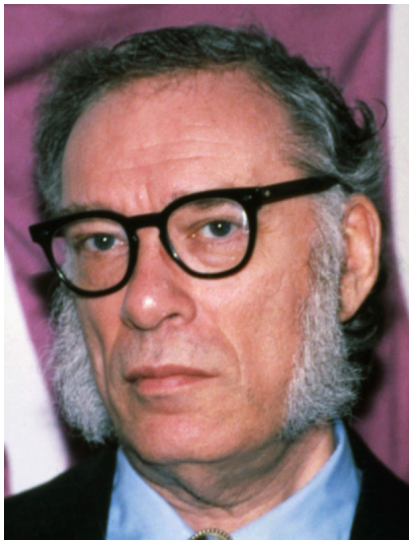
- Bill Gates (2015)
  - “I am in the camp that is concerned about super intelligence.”

- Elon Musk (2017)
  - “AI is a fundamental risk to the existence of human civilisation.”

- Henry Kissinger (2018)
  - “… whose culmination is a world relying on machines ungoverned by ethical or philosophical norms.”

Asimov's Three Laws

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.



Isaac Asimov
1920-1992

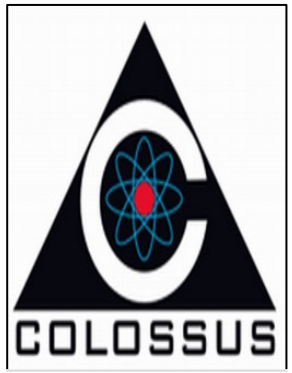
Google's Five Problems

- Avoid Negative Side Effects
  - How can we ensure that an AI system will not disturb its environment in negative ways while pursuing its goals?
- Avoid Reward Hacking
  - How can we avoid gaming of the reward function?
- Scalable Oversight
  - How can we efficiently ensure that a given AI system respects aspects of the objective that are too expensive to be frequently evaluated during training?
- Safe Exploration
  - How do we ensure that an AI system doesn't make exploratory moves with very negative repercussions?
- Robustness to Distributional Shift
  - How do we ensure that an AI system recognizes, and behaves robustly, when it's in an environment very different from its training environment?

research.googleblog.com/2016/06/bringing-precision-to-ai-safety.html

Ethical Issues

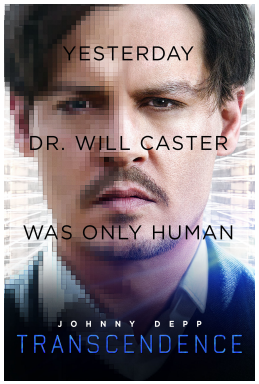
- End of human race
  - An unchecked AI system makes a mistake
  - Utility function has undesired consequences
  - Learning leads to undesired behavior
  - Singularity
- Friendly AI



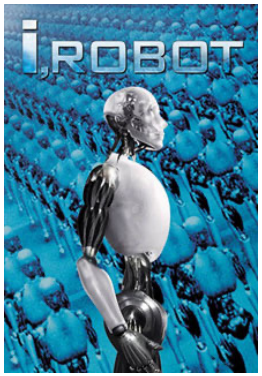
"Colossus: The Forbin Project" (1970)

"The Matrix" (1999)

"Terminator 3: Rise of the Machines" (2003)



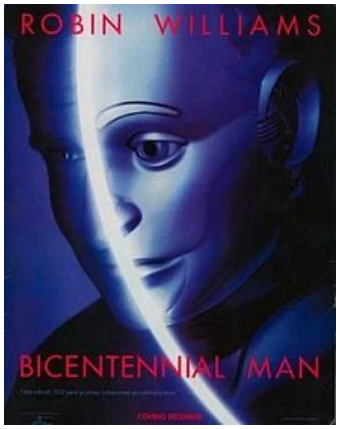
"Transcendence" (2014)



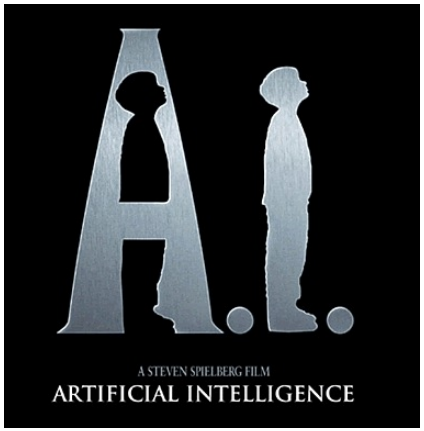
"I, Robot" (2004)

Ethical Issues

- Robot/AI rights



"Bicentennial Man" (1999)



"A.I. Artificial Intelligence" (2001)

"The Machine" (2013)

"Ex Machina" (2015)

Political Issues

- Artificial Intelligence for the American People
  - www.whitehouse.gov/ai/

Ethics

- Weak AI vs. Strong AI
- Controlling AI
- AI Laws
- AI Rights
- Human Future
- Policy