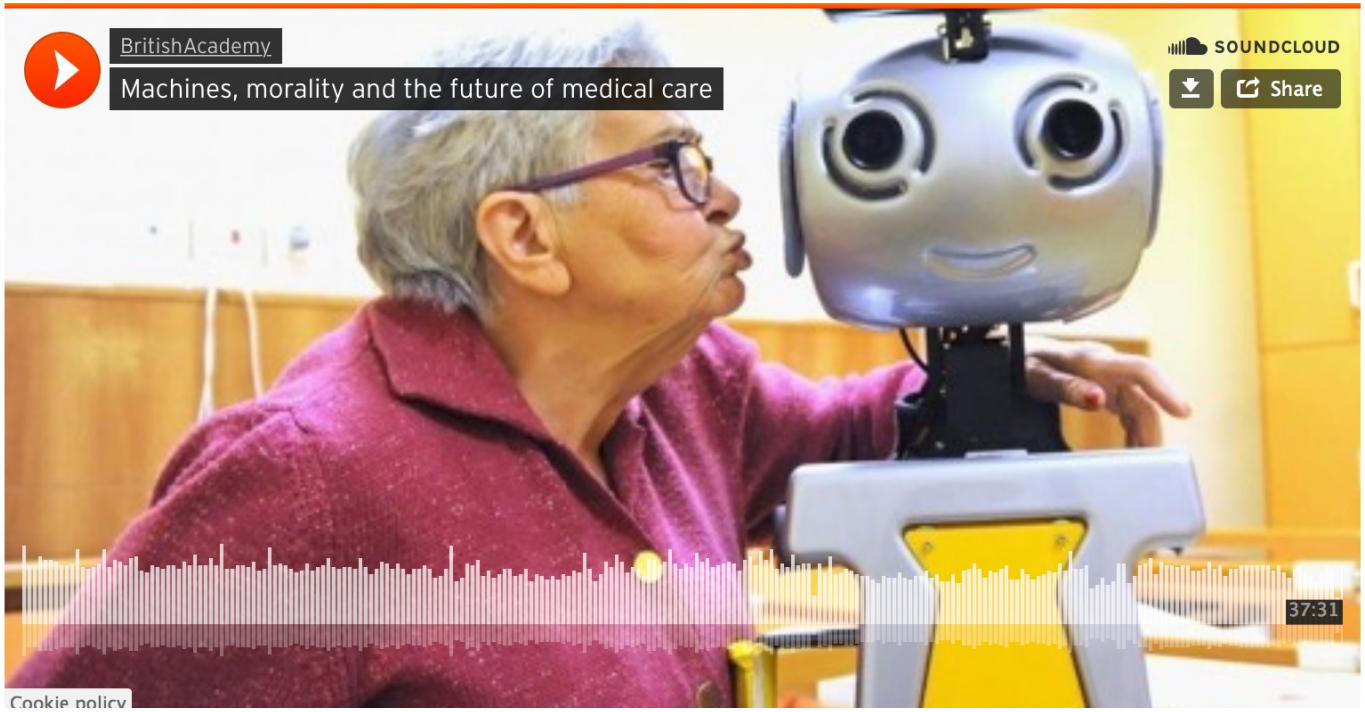
CS440/ECE448 Lecture 39: Ethics of Autonomous AI



Mark Hasegawa-Johnson, 4/2022,

With some slides by Svetlana Lazebnik

Image source: https://www.britac.ac.uk/audio/machines-morality-and-future-medical-care

Outline

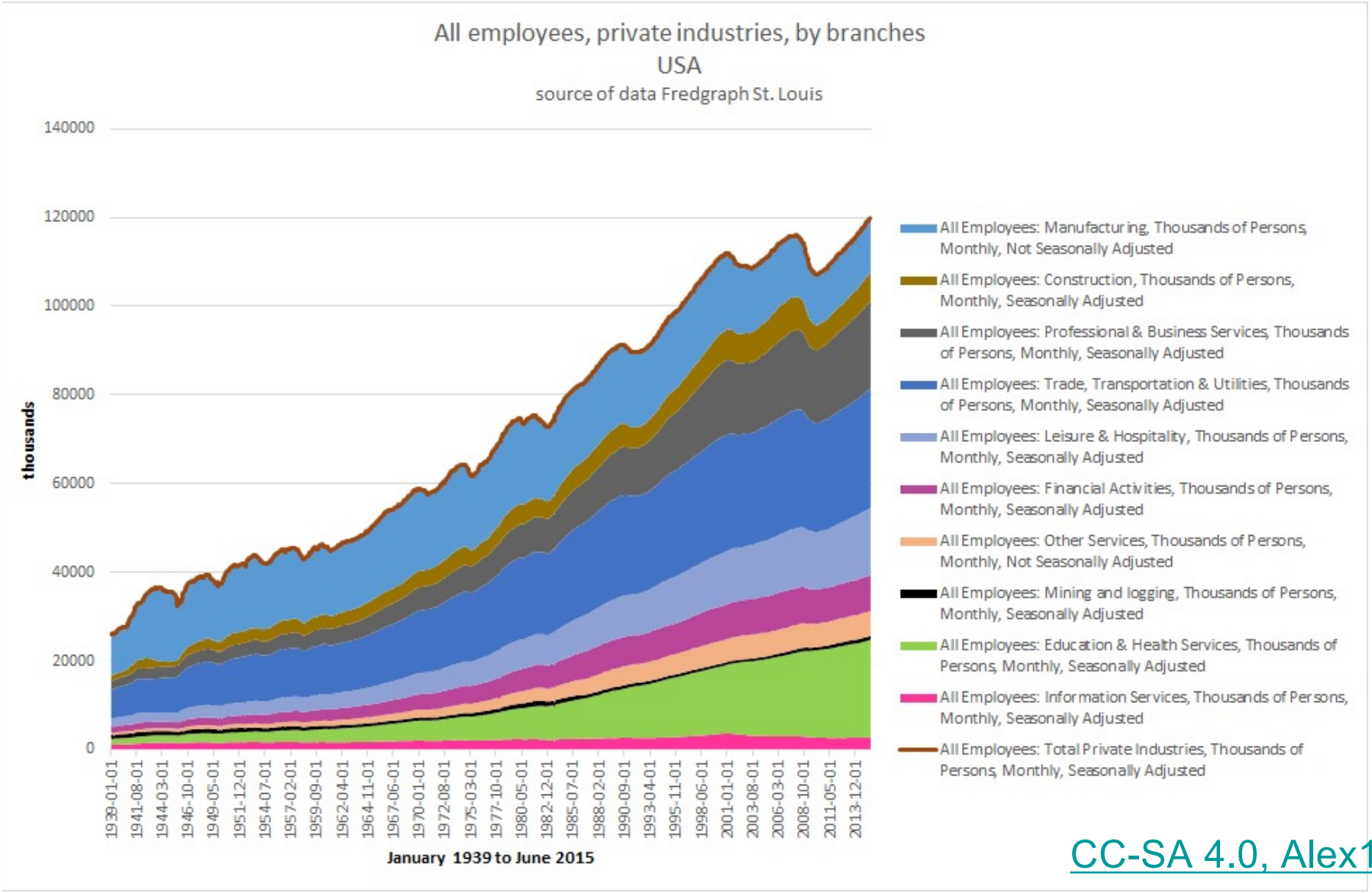
- Jobs
- Safety
- AI weapons
- Superintelligence
- Robot rights

AI and jobs

Source

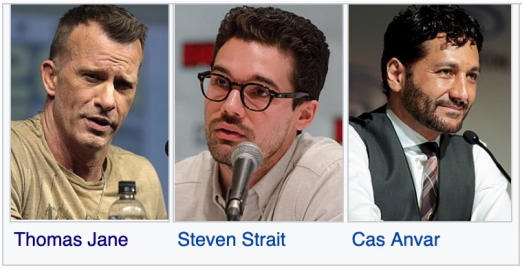
AI and jobs

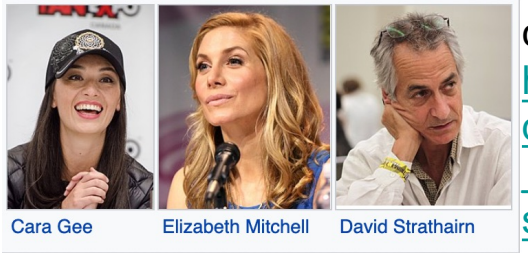
- Why we should worry
  - Oxford report: 47% of American jobs at high risk of automation in the next two decades
  - In the past couple of decades, manufacturing employment has dropped even as output kept rising; labor force participation among working-age males has been dropping
  - Truck driver is the most common job in over half the states
- Why we shouldn't worry
  - Historically, automation has destroyed jobs but added more new jobs



An alternative to work?

- In the series of novels *The Expanse*, authors Daniel Abraham and Ty Franck predict that most people will live on “basic income” provided by the government.
- People who want to earn more, by working, must first go to college.
- People who want to go to college must first work for a trial year in an uneducated job, to prove that they have the will to work.



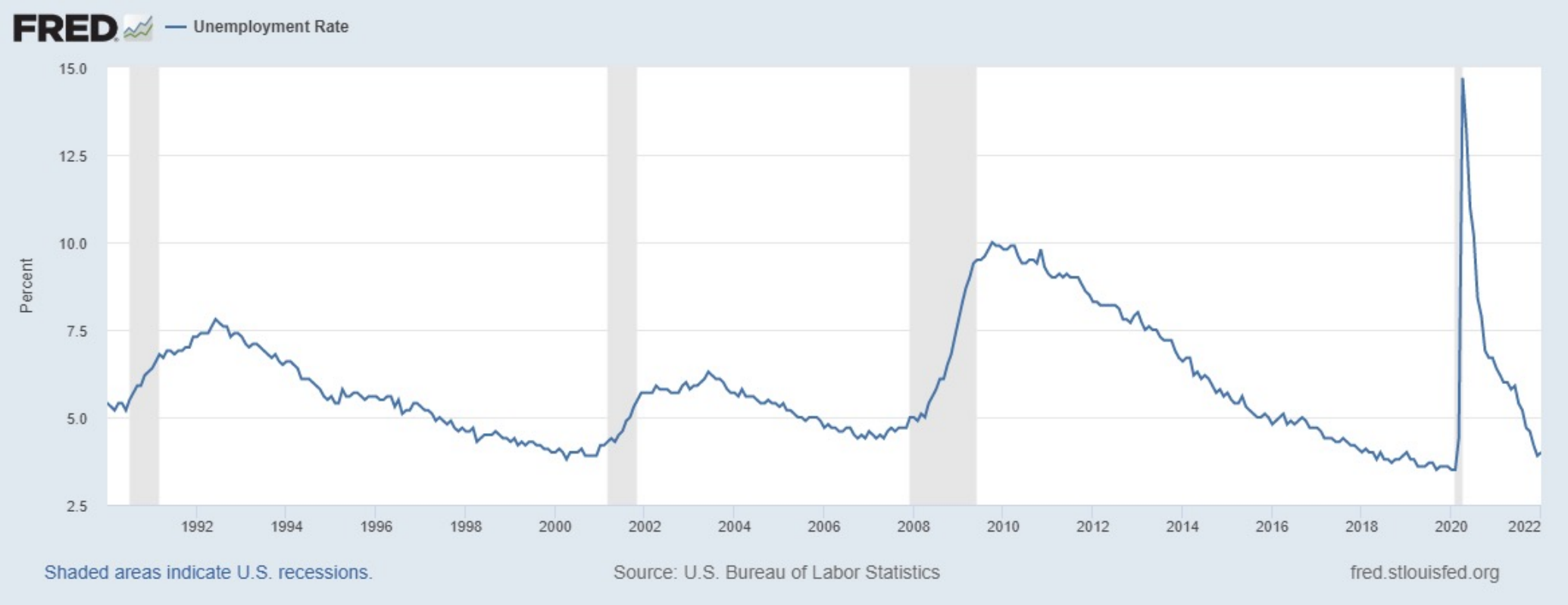


Actors from The Expanse TV series. Images contributed CC-SA by many contributors: https://en.wikipedia.org/wiki/The_Expanse_(TV_series)

Is “Basic Income” going to happen?

“Netherlands is among the nations with the shortest work weeks, according to *CNN Money*. Netherlands has an average of 29 working hours per week and an average annual income amounting to $47,000.” – https://nltimes.nl/2013/07/11/worlds-shortest-work-weeks

Is “Basic Income” going to happen in the United States?



US Unemployment rate, January 1990 to June 2021. Public domain image, https://fred.stlouisfed.org/graph/?g=FfBh

AI and jobs

- Reading
  - Moshe Vardi talk (YouTube)
  - AI NOW report (2017)
  - Technological unemployment (Wikipedia)
  - A world without work (The Atlantic, July 2015)
  - The automation paradox (The Atlantic, Jan. 2016)
  - AI will transform the economy. But how much, how soon? (New York Times, Nov. 2017)
  - Welcoming our new robot overlords (New Yorker, Oct. 2017)
  - The great tech panic: robots won't take all our work (Wired, Aug. 2017)

Outline

- Jobs
- Safety
- AI weapons
- Superintelligence
- Robot rights

AI safety

- Robustness to changes in data distribution
- Avoiding catastrophic “corner cases”
- Robustness to adversarial examples or attacks
- Avoiding negative side effects in reward function
- Avoiding “reward hacking”

- Reading: Concrete AI safety problems

Is autonomous driving more or less dangerous than human driving?

- Tesla (both human and autopilot): 1 accident per 4.34 million miles of driving
- All US (mostly human drivers): 3.26 trillion miles, 6.76 million accidents reported to police in 2019 = 1 reported accident per 482 million miles driven
- https://www.police1.com/patrol-video/articles/video-tesla-on-autopilot-slams-into-police-car-swipes-2-leos-0jqkJUHqiqVWxIhH/

Is autonomous driving more or less dangerous than human driving?

- Tesla:
  - 1 billion miles on autopilot, with 12 deaths reported (https://www.tesladeaths.com/)
- Record for human drivers is slightly better, depending on state laws
  - In the U.S.: 1.11 deaths per 100 million miles
  - In Illinois: 0.94 deaths per 100 million miles
  - https://www.iihs.org/topics/fatality-statistics/detail/state-by-state

Outline

- Jobs
- Safety
- AI weapons
- Superintelligence
- Robot rights

AI weapons

https://www.youtube.com/watch?v=aLS3Jljy1VA

Are autonomous weapons ethically acceptable?

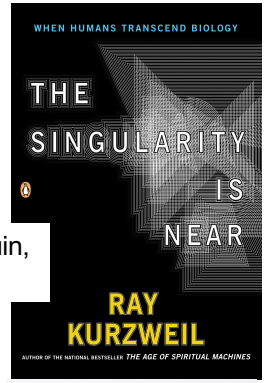
- The case in favor, e.g., https://blog.apaonline.org/2022/02/28/the-moral-case-for-the-development-of-autonomous-weapon-systems/
  - Autonomous weapons reduce casualties for the side using them
  - Autonomous weapons can select targets more precisely, thereby reducing casualties for the other side as well
- The case against, e.g., https://www.hrw.org/report/2018/08/21/heed-call/moral-and-legal-imperative-ban-killer-robots
  - Autonomous weapons encourage totalitarianism by permitting a small number of people to suppress a large number
  - Autonomous weapons may malfunction
  - Autonomous weapons diffuse responsibility for death, thereby encouraging war

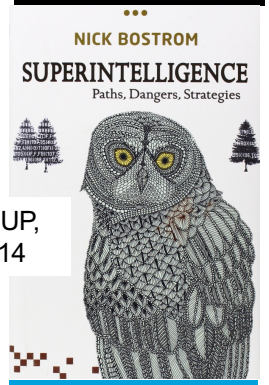
Outline

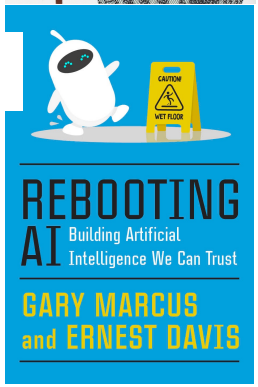
- Jobs
- Safety
- AI weapons
- Superintelligence
- Robot rights

Superintelligence: Brief history of a discourse conducted by best-selling authors

- The Singularity, Kurzweil, 2006:
  - Processing power of a chip will pass that of a human brain ~2030, and keep growing exponentially.
  - AI will become exponentially smarter than humans.
- Superintelligence, Bostrom, 2014:
  - That might not be good for humans.
- Rebooting AI, Marcus & Davis, 2019:
  - The problem with AI is not that it's too smart. The problem is that it's too dumb.

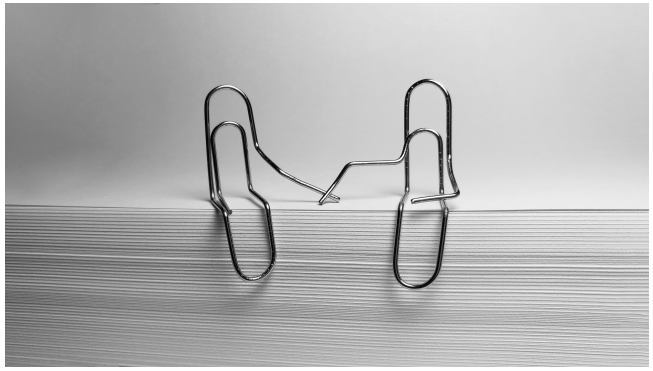






Superintelligence: The problem

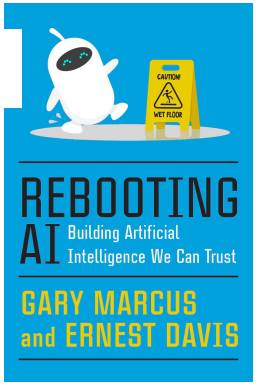
- Give an AI a very limited task (e.g., maximize the number of paper clips you can make)
- Give it sufficient intelligence that it can figure out how to take resources away from humans (e.g., it can figure out how to hack all existing computers, and how to build more computers)
- Result: all solid material on Earth, including all human bodies, are turned into paper clips



Superintelligence: The solution

- We need to understand what makes humans moral.
- We need to find a moral code that is binding on all intelligences, not just on human intelligences.
- This cannot be something that we build into the AI (like Asimov's "three laws of robotics"), because the AI will just figure out how to get around it. It needs to be an imperative that is naturally binding on all intelligences, regardless of whether they believe it or not.

The Fragility Argument



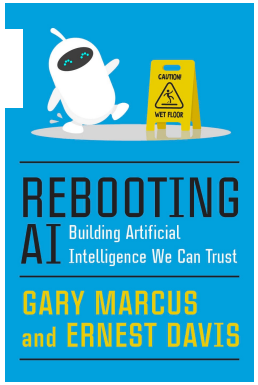
- Yes, Moore's law says that processing power doubles every two years, but…
- Linear increases in robustness (ability to devise solutions that are effective in stochastic real-world environments) require exponential increases in processing power.

$$IQ = \text{Robustness} = \log(\text{Procesing}) = \log(2^{0.5\times\text{year}})$$

…therefore, at best,

$$IQ = 0.5\times\text{year}$$

The Fragility Argument



The problem with AI is not that it's too smart. The problem is that it's too dumb.

- Privacy issues: the data is used in a way not permitted by its natural owners, not because the AI was too smart, but because the company used insecure passwords
- Bias and fairness issues: the AI doesn't understand the societal impact of its decisions
- Autonomous vehicle crashes: the AI makes mistakes

Outline

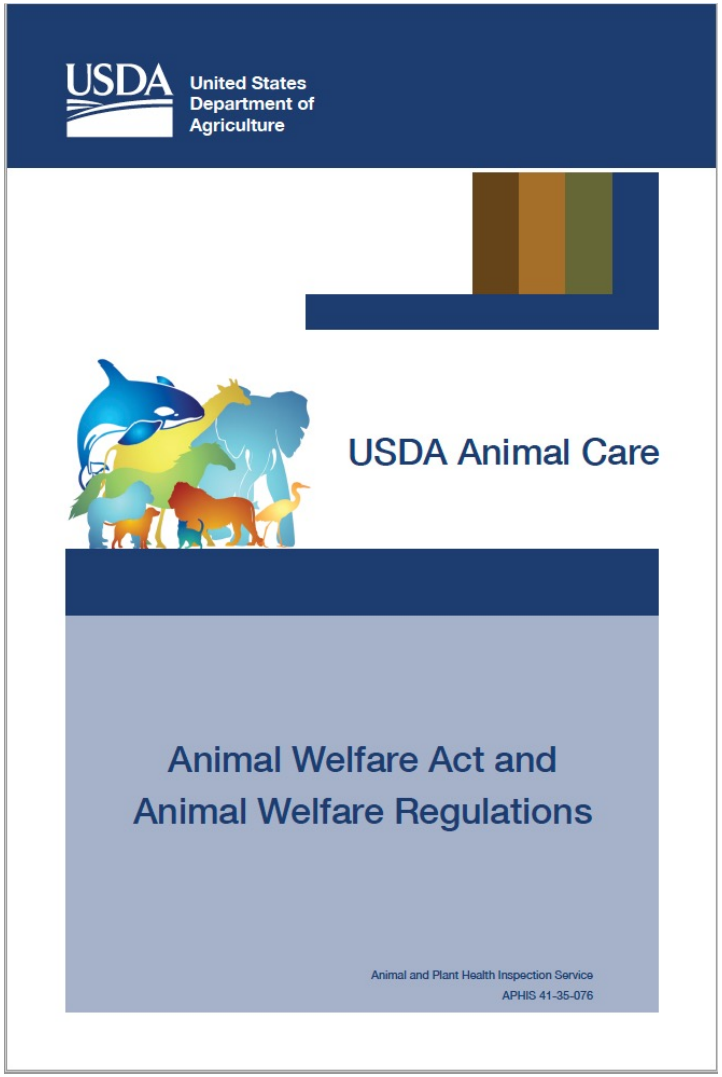
- Jobs
- Safety
- AI weapons
- Superintelligence
- Robot rights

Where do human rights come from?

- Aristotle: those with the ability to reason about the future have, therefore, the right to exercise their ability
- Aquinus: all those with souls have rights
- Locke: if rights are ignored by a government, humans can overthrow the government
- Finnis: the doctrine of rights helps us to construct a better society
- Nathwani: the doctrine of rights helps a government to maintain its legitimacy in the eyes of the governed

Do animals have rights?

The capabilities argument: creatures that can feel pain should not be subjected to avoidable pain

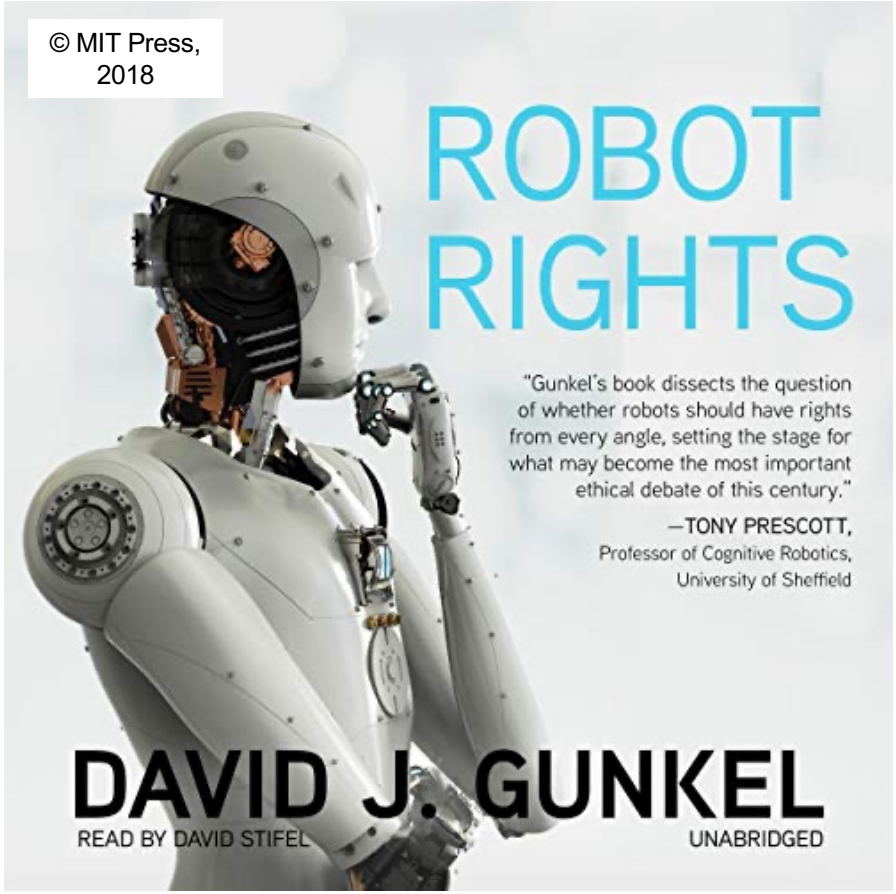


Do robots have rights?

Robots do not currently have

1. … the ability to rationally plan for an unlimited future, or
2. … the ability to feel pain.

But some day they may have those abilities (depending on how we define those abilities). In preparation for that future, maybe it's worth thinking carefully about the basis for human rights, to decide what would be the threshold cognitive ability that would require rights for robots.



The ethics of AI

- Jobs
- Safety
- AI weapons
- Superintelligence
- Robot rights