AI Ethics

4/28/17
Near-Term Ethical Concerns

AI displacing jobs

privacy

military AI

financial instability

DOW 9,869.62

998.50 / 9.2%
Discussion Questions

What are the redistributive consequences of AI? Are they fundamentally different from other technological innovations?

What new rules or norms regarding privacy will be required as AI becomes more pervasive?

Who should be blamed if an autonomous agent causes harm? Ethically? Legally?

What safeguards can/should AI developers put in place to ensure that the technology isn’t misused?
Longer-Term Ethical Concerns

- Making ethical decisions autonomously
- Single-task optimizers
- Singularity
Azimov’s Laws of Robotics

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.

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
Discussion Questions

How should autonomous vehicles solve the trolley problem? What about other ethical dilemmas?

How can we implement single-task agents to ensure they are safe?

Is achieving AI the same as the singularity? Should we fear the singularity or hope for it?

Is there such a thing as provably-safe true AI? What might that look like?