1 handout: slides
ethics for AI research
lethal autonomous weapons
trustworthy
transparency, legibility, explainable
value alignment
asst 12
project presentations: Wed May 8, 9-12
donuts, beverages
should AI be pursued?
  jobs, evil uses, danger
  accuracy, efficiency, convenience

how should AI be pursued
  objective: accuracy vs equal treatment
  data collection: accuracy vs privacy

should robots be moral agents?
  Bryson: no
  transhumanism
Lethal Autonomous Weapons Systems

examples

- mines
- Phalanx, Iron Dome
- cruise missiles
- Samsung sentry
- slaughterbots

ethics

- -: no human judgment, needs strong limitations
- +: no emotion, fatigue
fears

- success in limited tasks $\neq$ general intelligence
- intelligence $\neq$ desire to replace, destroy, dominate

mitigation

- software engineering
- formal verification
- transparency
- legibility
- explainability
Value Alignment

- hard to formalize
- no malicious intent necessary
  - genie in lamp, Sorcerer’s Apprentice, King Midas
- ‘instrumental goals’: ensure success of primary goal by preventing interference, acquiring resources (material and financial)
- how to prove alignment?
  - superintelligence inherently hard to predict
Trolley Problem


- head toward person wearing seatbelt or person not?
- push large person in front of trolley to save multiple people?
- your child or multiple old people?
- kill one person or give multiple permanent suffering?
- who is responsible?
- who certifies car?
Ideas for Learning Values

- direct programming
- observe humans
- debates between agents, scored by humans
- iterated amplification
- recursive reward modeling
1. among all possibilities, mind that respects human values is statistically improbable
   even getting close to human values might not be enough. also, could diverge over time
2. AGI likely to reach singularity (‘hard takeoff’)
3. aligned superintelligence likely dangerous

retort:
1. not obvious
   evolution implies likely or robust
2. soft take-off seems more likely
3. not clear. plenty of more pressing things to worry about
**Rules for Robots**

**Steve Omohundro:** scaffolding: each system proves safety of next
  or system shuts itself off if it can’t meet safety constraints

**Eliezer Yudkowsky:** coherent extrapolated volition: be what humans would want you to be
  human does not get to choose!

**David McAllester:** servant mission: within the law, fulfill my requests
  society makes laws
  avobot might have incentive to lie or restrict information access
The Off Switch Game (Hadfield-Menell et al, IJCAI-17)

- robot maximizes human’s utility
- human utility uncertain
- switching off provides useful information for future robot actions
- even high confidence that \( a \) is good for \( H \) allows off
What question didn’t you get to ask today?
What’s still confusing?
What would you like to hear more about?

Please write down your most pressing question about AI and put it in the box on your way out.

Thanks!