Threats of AI

Marek Behún
<kabel@blackhole.sk>

“Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans.”

Nick Bostrom, 2003

Yet another Doomsday prophecy?

- Who are Millenialists?
  - Montarists of Turkey in 2nd century
  - Anabaptists of Netherlands in 16th century
  - Sabatianists of Izmir in 17th century
  - Millerists of USA in 19th century
- We have a lot in common with Millennialists
- One of the reason why many people, even scientists, do not take this seriously
- … so why doing this?

What is ASI?

- AGI which is also a superintelligence
- Decisionmaker better in all decisionmaking a human can do
- Even the most gifted humans
- This leads directly to an intelligence explosion
- Which in turns very probably leads to paperclip maximizers

Humans are not Built to Think About AI

- For ordinary humans
  - Cognitive biases
    - Availability heuristics
    - Overestimation of probability of conjunctive events
    - Anchoring
    - Scope insensitivity
    - Optimism bias
    - …
  - Plain stupidity
- For scientists
  - Cognitive biases again
  - Rationalization
- Basically, humans are systematically insane

Rationalization? But...

- An ASI won’t come in at least 100 years!
  - Even if P(ASI in 100 years) = 10%, this is still catastrophic
  - Less than 10 revolutionary discoveries are needed
    - Working exocortex, for example, could lead to an ASI
- An ASI won’t want to harm us! It will do what we want it to do!
  - Value is Complex and Fragile
  - Paperclip maximizer
  - Even it this is true, there are other problems
- ...

Other Problems?

- Values of Human Civilization change over time
  - Some time ago slavery was accepted
  - Do we want to force our values on future, probably better humans?
- Mindcrime
  - Programs can be people too, therefore objects of ethical value
  - Church-Turing thesis
- ...

The Alien Metaphor

- If an alien spaceship came into the Solar System and trasmitted the message:

  “WE WILL COME IN 30 YEARS”

  - Will they come in peace? Are they good aliens?
  - Or are they preparing us for some sport war?
  - Everyone would freak out
  - We would get our shit together and start doing something (or maybe not)
  - ASI is similar

AI Research is too Open

- Imagine the laws of Physics were somehow so that everyone could create a device with the power of an Atomic Bomb from the garbage in their backyard
  - If you were the researcher who discovered these laws, would you openly share them?
  - Basically, we are preparing Computer Programming to be like this – sooner or later, it will be possible to create a potential ASI with a crazy utility function from stuff found on the internet
    - What will happen if a sufficiently smart idiot finds the stuff?

Conclusion

- ASI is potentially very dangerous
- And also possible within “near” future
- And yet very little is being done to prevent future idiots to play with it
  - By idiots we mean people untrained in rational judgement or those who often fall prey to cognitive biases or rationalization… basically almost everyone

- Is at least something done?
  - Future of Humanity Institute
  - Machine Intelligence Research Institute

Literature

- www.intelligenceexplosion.com
- www.lesswrong.com
- www.arbital.com
- Judgement under Uncertainty: Heuristics and biases (Amos Tversky and Daniel Kahneman)
- Influence: Science and Practice (Robert Cialdini)
- Wikipedia
- Reddit :)