Review on AGI Safety Research

Turku AI Society

Juho Vaiste, 15.5.2018

Understanding AGI

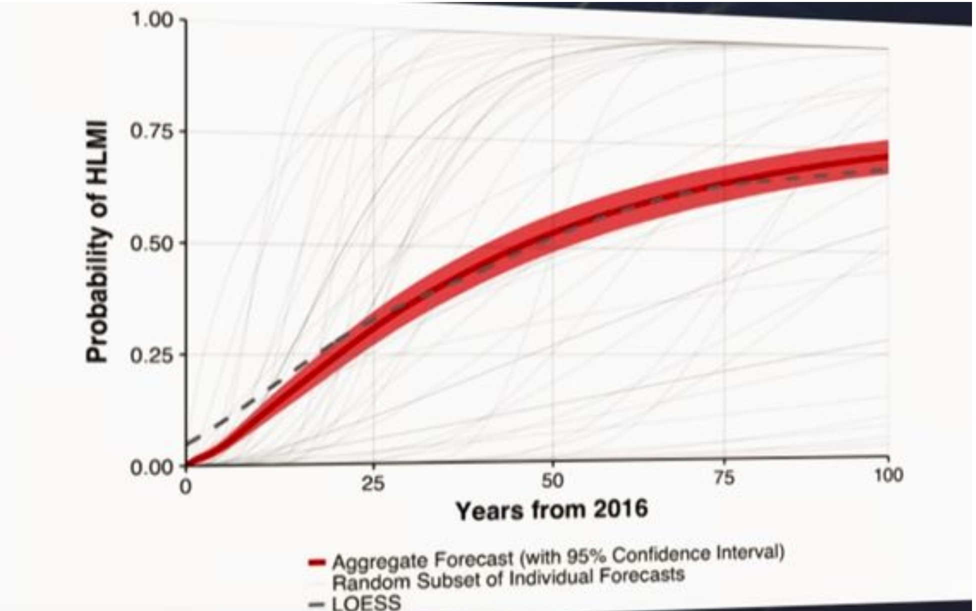
- Defining intelligence: “Intelligence measures an agent’s ability to achieve goals in a wide range of environments.”

- Orthogonality thesis
- Instrumental goals

- Formalizing AGI (AIXI) and alternate views (not an agent)

Predicting AGI Development

- Estimations done also by others than Bostrom. Not specific change during the last 5 years?



Predicting AGI Development

- Estimations vary a lot
- One perspective for a more in-depth look:
    - Chalmers’s analysis (2010) vs Dreyfus, Lucas, Penrose, Searle

Predicting AGI Development: Singularity?

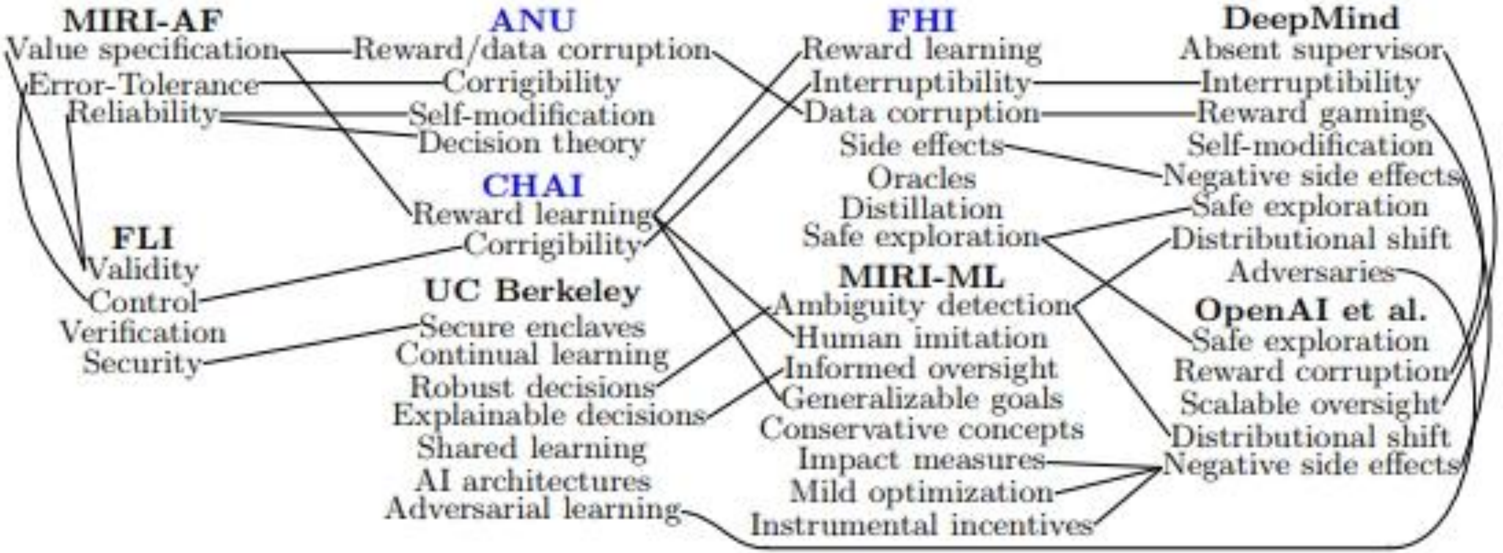
- Self-improvement as a instrumental goal → recursive self-improvement → intelligence explosion (Bostrom, 2014; Good, 1966; Hutter, 2012a; Kurzweil, 2005; Vinge, 1993; Yudkowsky, 2008b)

Predicting AGI Development: Singularity?

- Against (Walsh, 2016)

  - How to measure intelligence?
  - How much smarter is faster thinking?
  - Why human intelligence level have some kind of special meaning?

Institutes on AGI Safety



Problems with AGI

- How do we get an AGI to work towards the right goals?
- How can we make an agent that keeps pursuing the goals we have designed it with?
- If we get something wrong in the design or construction of an agent, will the agent cooperate in us trying to fix it?
- How to design AGIs that are robust to adversaries and adversarial environments?
- Safe learning: AGIs should avoid making fatal mistakes during the learning phase.
- How can we build agent's whose decisions we can understand?
- Societal consequences: AGI will have substantial legal, economic, political, and military consequences.

Summary

- **A small but emerging field**

  Still relatively easy to map the earlier literature; not gonna be the case in a couple of years

- **Moving from philosophy to mathematical fields**

- **Important**

  Hope to see a few dozen AGI researchers in Finland

Let’s list AGI researchers in Finland

- Kaj Sotala
- Harri Valpola and Curious AI team