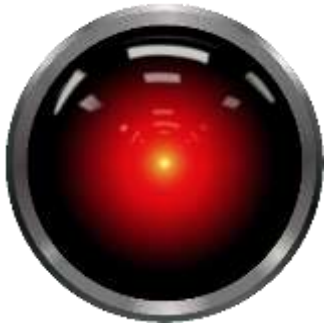
Safe Baby AGI

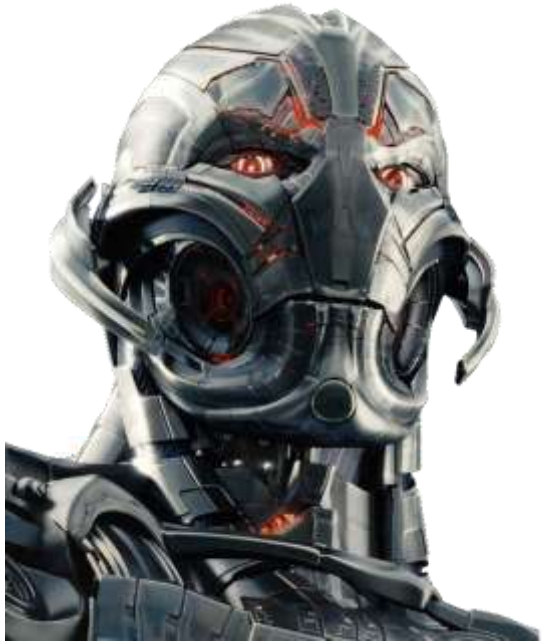
Jordi Bieger, Kristinn R. Thórisson & Pei Wang
School of Computer Science | Center for Analysis and Design of Intelligent Agents

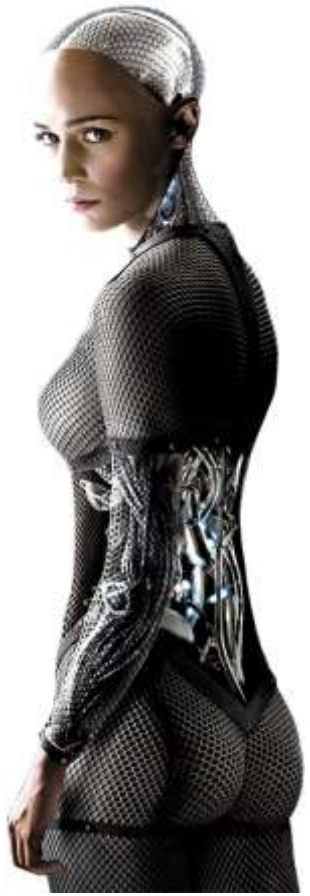
Introduction

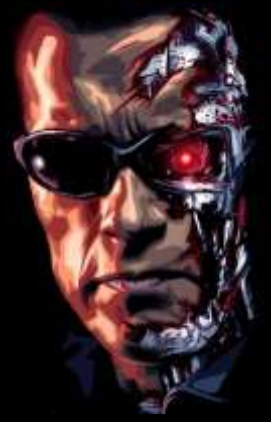
- Baby AGI
- Nature *and* nurture
- Intelligence and knowledge
- Friendliness

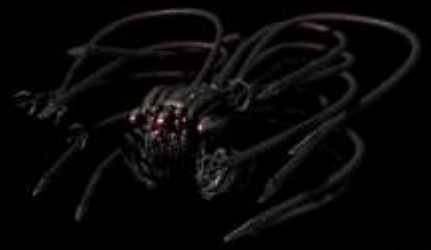
Rise of the Machines











Bigger, Stronger, Faster

- Assuming that the AI is willing, would it be able?
- Hard takeoff
  - Speed (hardware overhang)
  - Number (collective intelligence)
  - Quality (reprogramming and self-programming)
  - Content (availability of digital knowledge)

Appetite for Destruction

- Would the AI be willing to destroy humanity?
  - No love or hate for humans by default
  - Instrumental convergence / basic AI drives
    - Survival
    - Self-improvement
    - Resource acquisition
  - Difficult to provide human values

Institutions

- Machine Intelligence Research Institute (MIRI)
- Future of Humanity Institute (FHI)
- Future of Life Institute (FLI)
- Centre for the Study of Existential Risk (CSER)
- Global Challenges Foundation (GCF)
- Lifeboat Foundation
- …

Coping Strategies

- Confinement
- Monitoring
- Design 100% safe AGI
- Regulation

Bounded & Adaptive

- Knowledge and resources are bounded
- Knowledge and resources change continuously
  - Unpredictable in complex environments
- Behavior changes as new knowledge and skills are acquired and the AI adapts to the environment
- Completely general intelligence is unpredictable
- No hard guarantees
- Nature and nurture
  - Nature: design, implementation, innate qualities
  - Nurture: experience, interaction, acquired qualities
- Experimentation and the scientific method

Baby AGI

- Design and implementation are like an embryonic period
- Switching the system on is a kind of birth event

- Baby AGI lacks knowledge and skills
- Learns from interaction with the environment, which may include us as teachers
- Neither willing nor able to destroy humanity
- No basic AI drives (possibly)

Limiting Baby AGI

- Confinement and monitoring baby AGI should be simpler
- Such a system is not yet smart enough to deceive
- If becoming smart and powerful enough to escape, requires the AI to escape, it won't happen

Intelligence and Knowledge

- Particular knowledge is needed for particular realizations, even with high intelligence
- Any real system will have gaps in its knowledge
- No reason to actively seek out this particular knowledge
- Focus on the task at hand

Required Insights

- “super powerful AI should exterminate humans”
- “I can become this powerful”
- “this is a good idea”
- “I need to make a plan”
- “I need to hide my intentions, knowledge and capabilities”

- Long chains of “coincidental” insights, unlikely in lab

Risky Business

- Cost vs. (unknown) benefit
- Humans will kill the AI for even contemplating this
- More powerful AI may also be “laying in wait”
- Known fallibility

Nurturing Beneficial AGI

- Opportunity and responsibility to guide our AIs
- Emphasize effective and peaceful ways to accomplish goals
- Teach the risks of aggression and value of friendship
- Moral stages of development
- Value learning

Conclusion

- Baby AGI can be adequately controlled
- We must consider nature *and* nurture
- Experimentation on real AGI systems might provide insights into capability *and* safety for future generations
- Humans use technology for good or bad