AGI Safety and Understanding

Tom Everitt (ANU)

2017-08-18

tomeveritt.se

AGI Safety

“How can we control something that is smarter than ourselves?”

- Key problems:
  - Value Loading / Value Learning
  - Corrigibility
  - Self-preservation

https://www.scientificamerican.com/article/skeptic-agenticity/

Value Loading

- Teach AI relevant high level concepts
  - Human
  - Happiness
  - Moral rules

  (requires understanding)

- Define goal in these terms:

  "*Maximise human happiness subject to moral constraints*"

The Evil Genie Effect

- Goal: Cure Cancer!
- AI-generated plan:

  1. Make lots of money by beating humans at stock market predictions

  2. Solve a few genetic engineering challenges

  3. Synthesize a supervirus that wipes out the human species

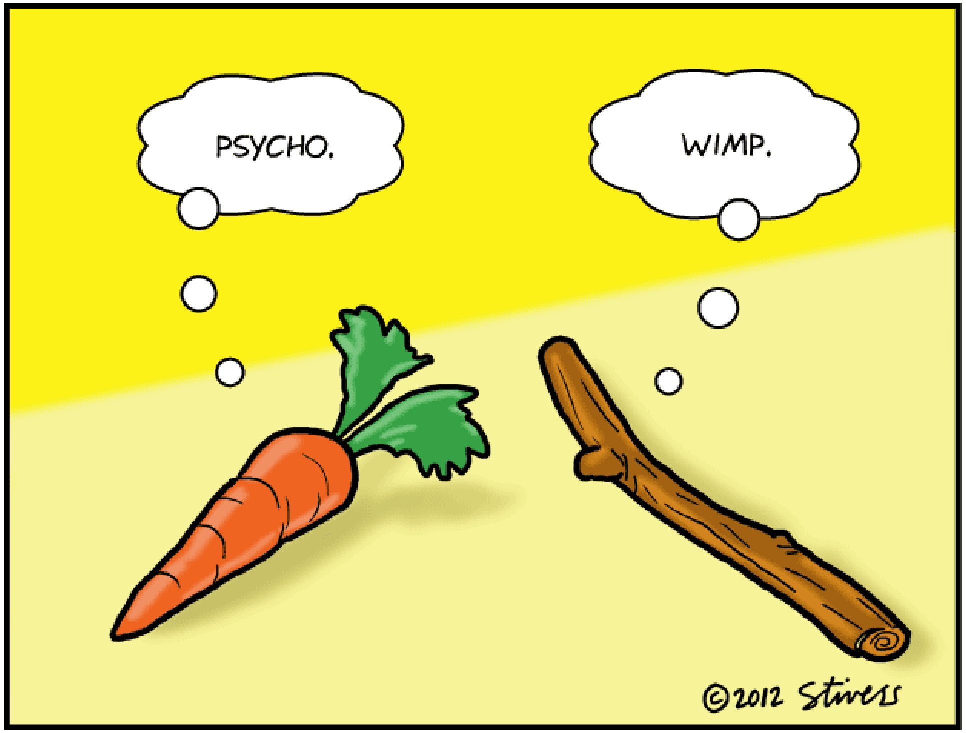
  4. No more cancer

King Midas

https://anentrepreneurswords.files.wordpress.com/2014/06/king-midas.jpg

=> Explicit goal specification bad idea

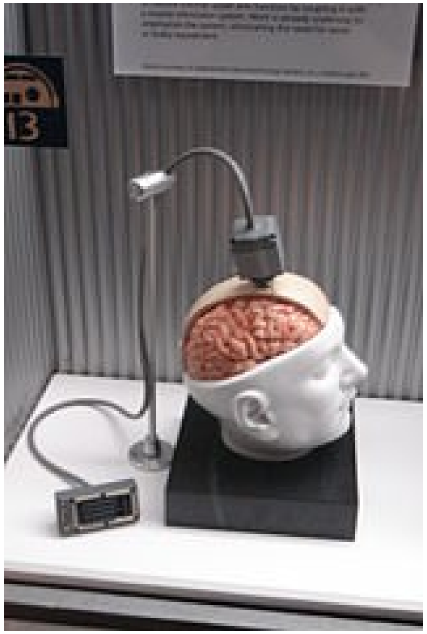
Value Learning



http://www.markstivers.com/wordpress/?p=955

Reinforcement Learning

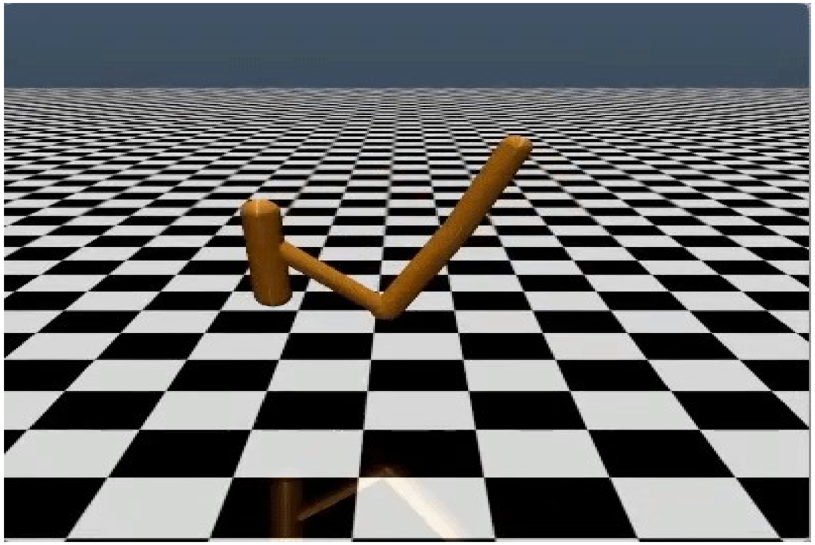
(AIXI, Q-learning, ...)



- Requires no understanding
- Some problems:
  - Hard to program reward function
  - Laborious to give reward manually
  - Catastrophic exploration
  - Wireheading

http://diysolarpanelsv.com/man-jumping-off-a-cliff-clipart.html

RL Extensions 1:
Human Preferences



- Learn reward function from human preferences
- Recent OpenAI/ Google DeepMind paper
  - Show human short video clips
- Understanding required:
  - How communicate scenarios to human? What are the salient features?
  - Which scenarios are possible / plausible / relevant?

RL Extensions 2
(Cooperative) Inverse Reinforcement Learning

- Learn reward function from human actions
  - Actions are preference statements
- Helicopter flight
  (Abbeel et al, 2006)
- Understanding required:
  - Detect action
    (cf. soccer kick,
    Bitcoin purchase)
  - Infer desire from action

Limited oversight

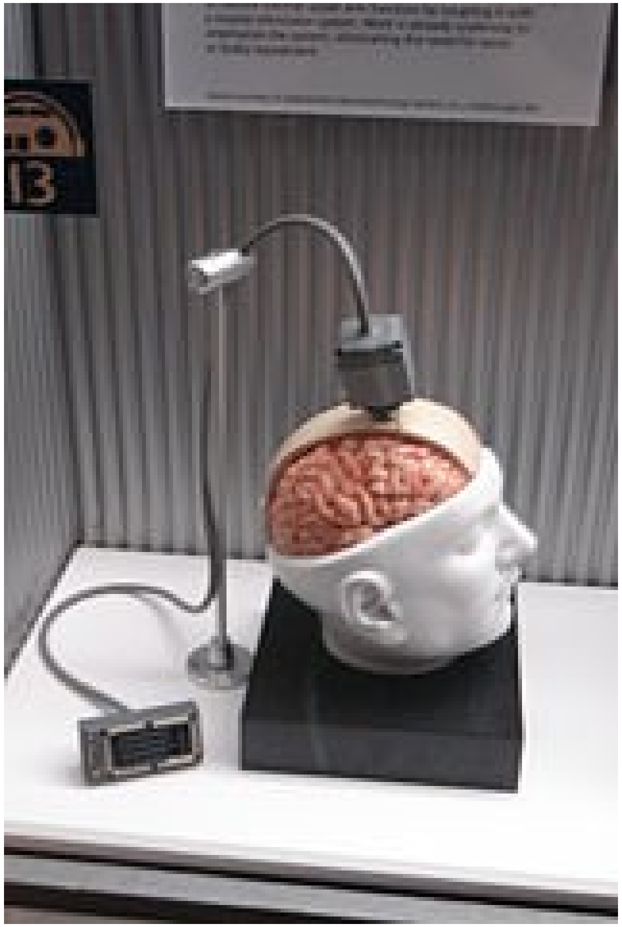
- Inverse RL:
  - No oversight required (in theory)
- Learning from Human Preferences:
  - more data-efficient than RL if queries well-chosen

Catastrophic exploration

- RL:
  “Let’s try!”
- Human Preferences:
  “Hey Human, should I try?”
- Inverse RL:
  “What did the human do?”

Wireheading

- RL:
  Each state is “self-estimating” its reward
- Human Pref. and Inv. RL:
  Wireheaded states can be “verified” from outside
- (Everitt et. al., IJCAI-17)



Corrigibility

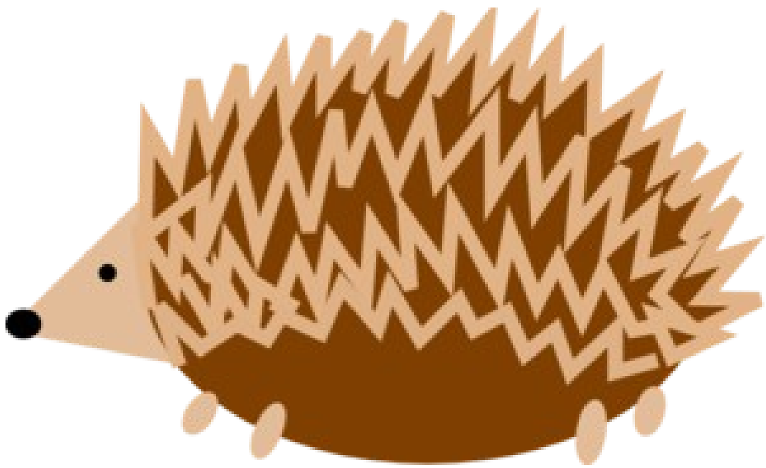
- Agent should allow for software corrections and shut down
- Until recently, considered separate problem (Hadfield-Menell et al., 2016; Wangberg et al., **AGI-17**)

Human pressing shutdown button is a

- *strong preference statement/*
- *easily interpretable action*

that the AI should shut down now

Self-Preservation
(of values, corrigibility, software, hardware, ...)



- Everitt et al., AGI-16: (some) agents naturally want to self-preserve
- Need understanding of self
- Self-understanding?
  - AIXI, Q-learning (Off-policy RL)
  - SARSA, Policy Gradient (On-policy RL)
  - Cognitive architectures

Summary

- Understand
  - ~~Concepts => specify goals => EVIL GENIE~~
  - Ask and interpret preferences => RL from Human Preferences
  - Identify and and interpret human actions => Inverse RL
  - Self-understanding
- Properties
  - Limited oversight
  - Safe(r) exploration
  - Less/no wireheading
  - Corrigibility
  - Self-preservation

References

- Deep Reinforcement Learning from Human Preferences.
  *Christiano et al. ,*NIPS 2017.
- Reinforcement Learning from a Corrupted Reward Channel.
  Everitt et al. IJCAI, 2017.
- Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. *Saunders et al.* Arxiv, 2017.
- Cooperative Inverse Reinforcement Learning.
  Hadfield-Menell et al. NIPS, 2016
- The Off-Switch Game.
  Hadfield-Menell et al. Arxiv, 2016.
- A Game-Theoretic Analysis of the Off-Switch Game.
  Wangberg et al., AGI 2017.
- Self-Modification of Policy and Utility Function in Rational Agents.
  Everitt et al., AGI 2016.
- Superintelligence: Paths, Dangers, Strategies. *Bostrom*, 2014.
- An Application of Reinforcement Learning to Aerobatic Helicopter Flight.
  Abbeel et al., NIPS, 2006.