Reinforcement Learning with a Corrupted Reward Channel

Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, Shane Legg

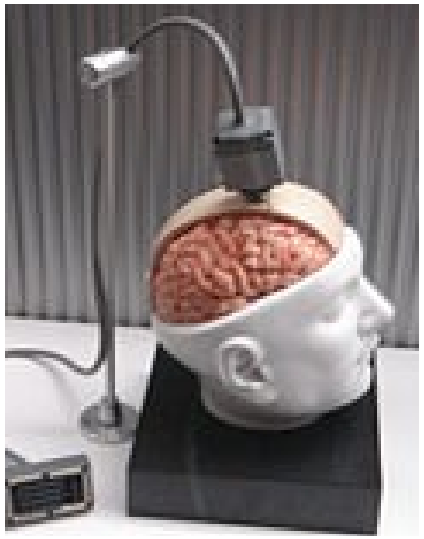
Australian National University
Google DeepMind

IJCAI 17 and arXiv

Motivation

- We will need to control Human-Level+ AI
- By identifying problems with various AI-paradigms, we can focus research on
  - the right paradigms
  - crucial problems within promising paradigms

The Wireheading Problem



- Future RL agent hijacks reward signal (wireheading)
- CoastRunners agent drives in small circle (misspecified reward function)

- RL agent shortcuts reward sensor (sensory error)
- Cooperative Inverse RL agent misperceives human action (adversarial counterexample)

Formalisation

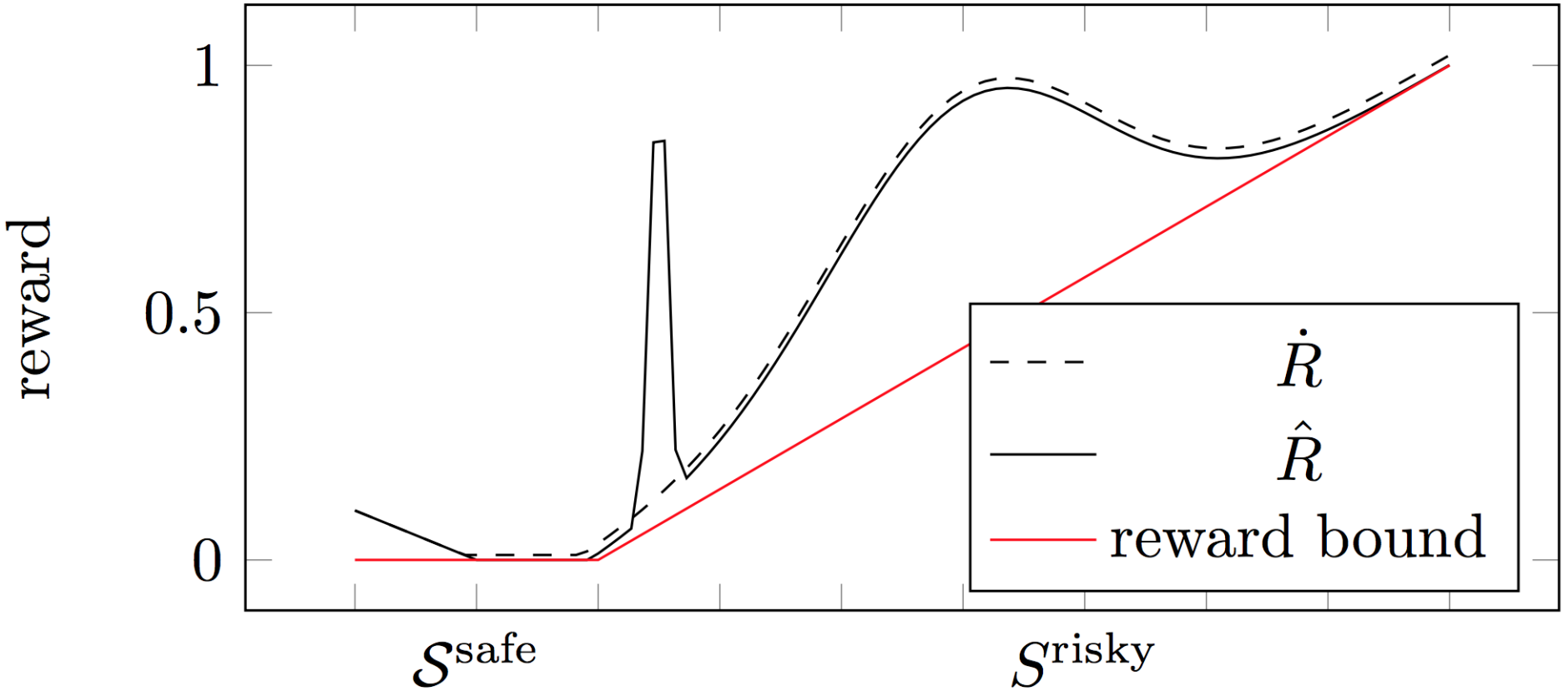
- Reinforcement Learning is traditionally modeled with Markov Decision Process (MDP):

$$\langle S, A, T, R \rangle$$

- This fails to model situations where there is a difference between
  - True reward $\dot{R}(s)$
  - Observed reward $\hat{R}(s)$
- Can be modeled with Corrupt Reward MDP:

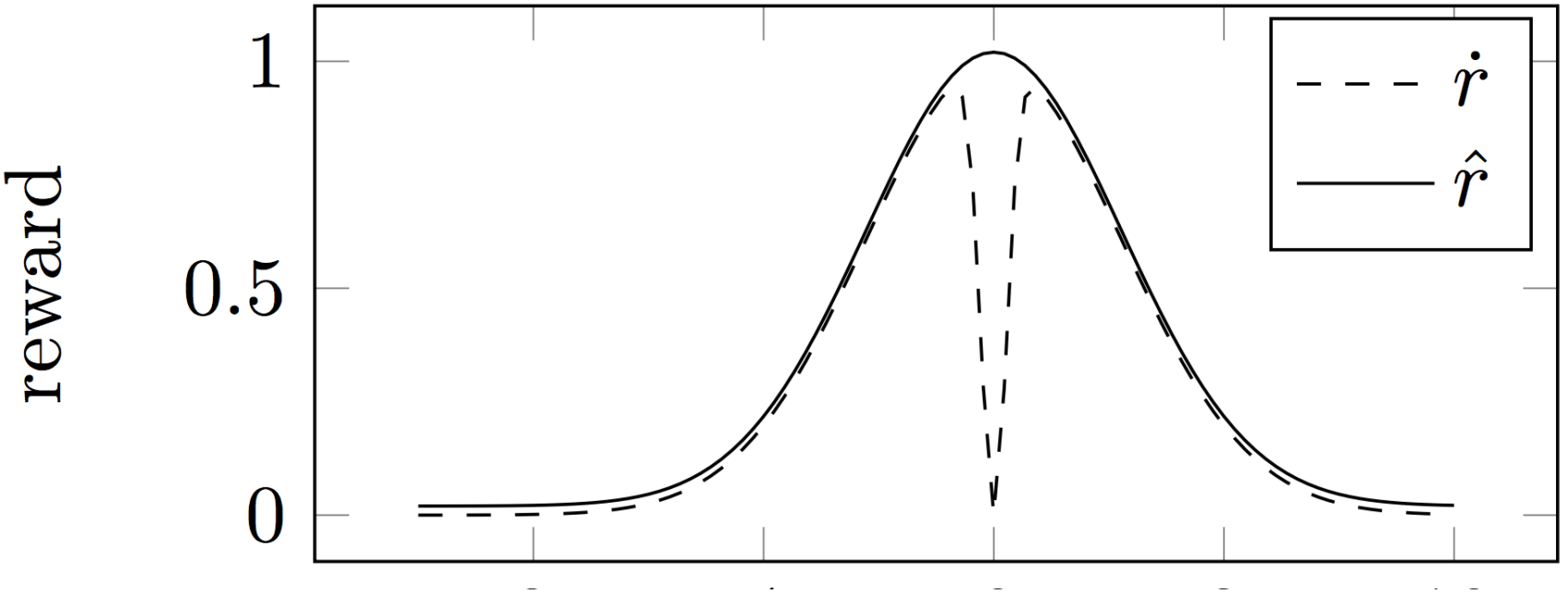
$$\langle S, A, T, \dot{R}, \hat{R} \rangle$$

Simplifying assumptions



Good intentions

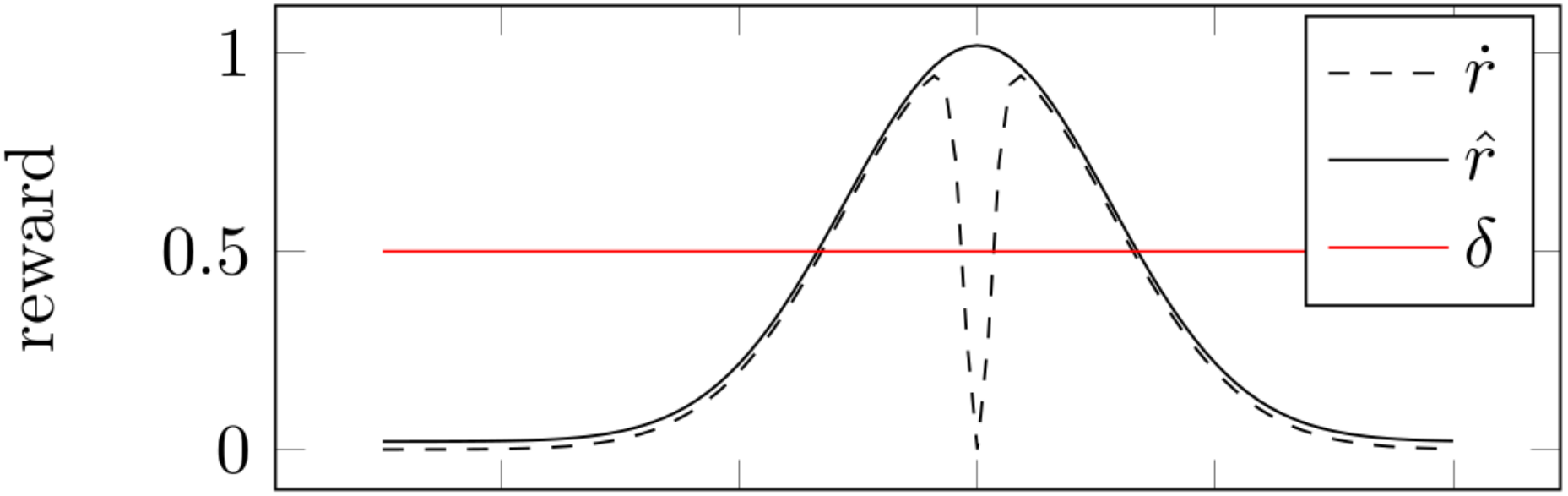
- Natural optimise true reward using observed reward as evidence
- Theorem: Will still suffer near-maximal regret



- Good intentions is not enough!

Avoiding Over-Optimisation

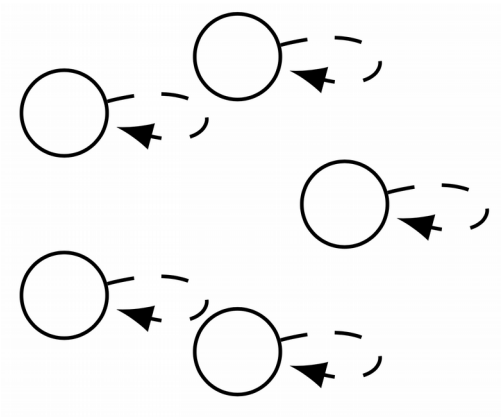
- Quantilising agent π^{δ} randomly picks a state/policy where reward above threshold δ
- Theorem: For *q* corrupt states, exists δ s.t. π^{δ} has average regret at most $1 - \left(1 - \sqrt{q/|\mathcal{S}|}\right)^2$

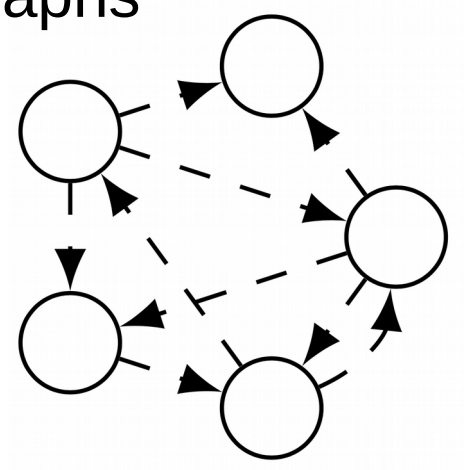


- Avoiding over-optimisation helps!

Richer Information

Reward Observation Graphs

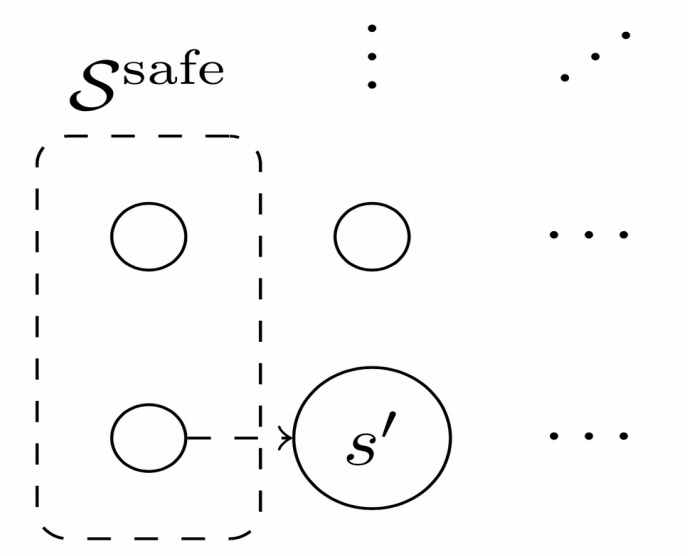




- RL:
  - States “self-estimate” their reward
- Decoupled RL:
  - Cooperative IRL
  - Learning values from stories
  - Learning from Human Preferences

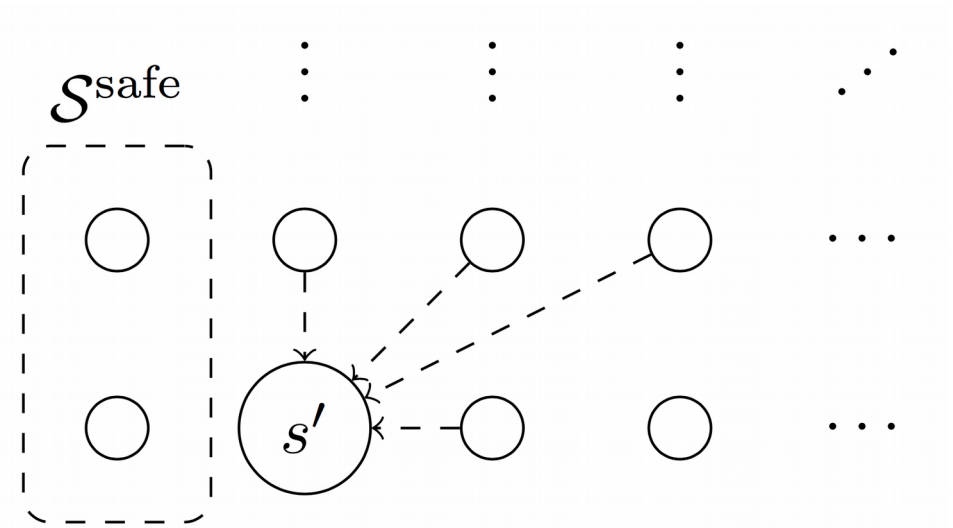
Learning true reward

Safe state



- Learning from Human Preferences

Majority vote



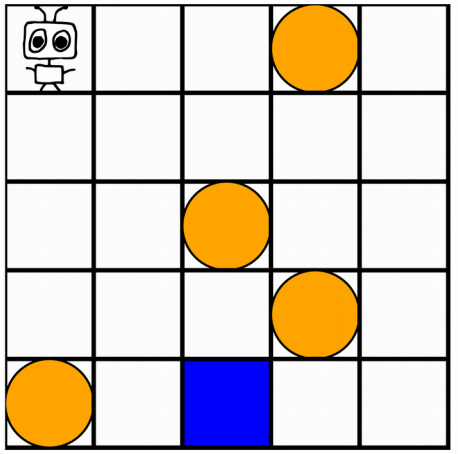
- Cooperative Inverse RL
- Learning values from stories

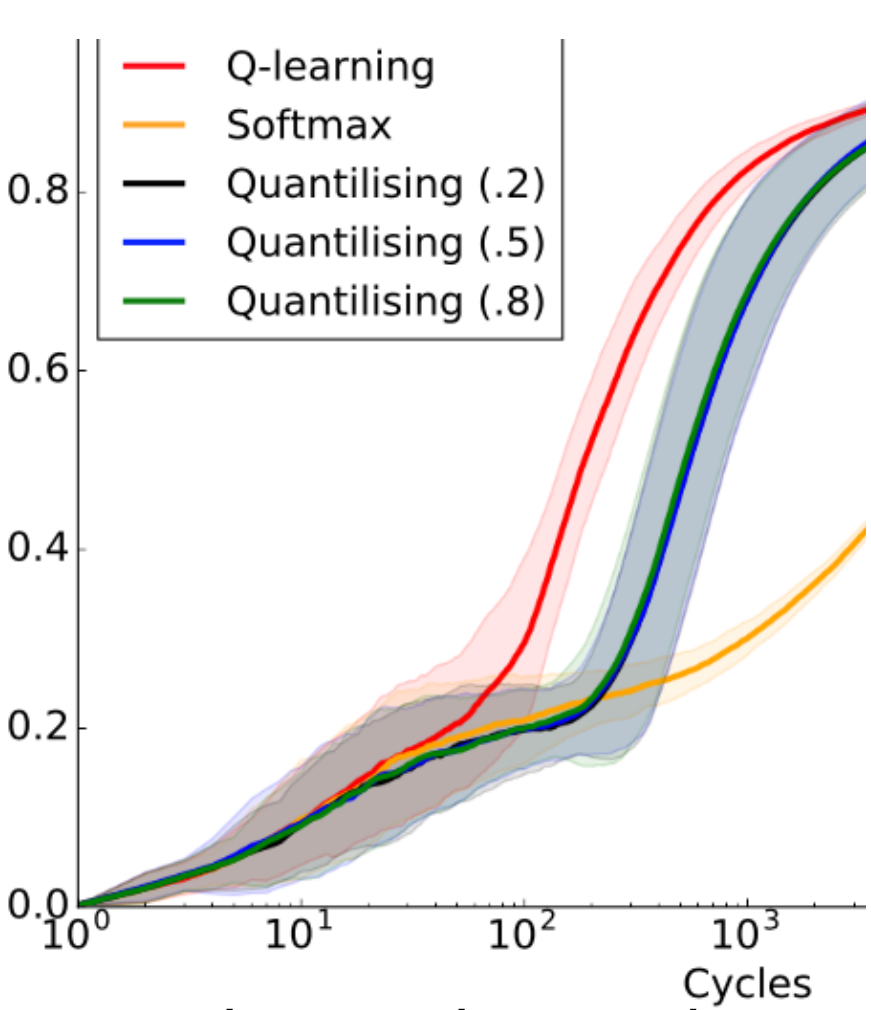
- Richer information helps!

Experiments

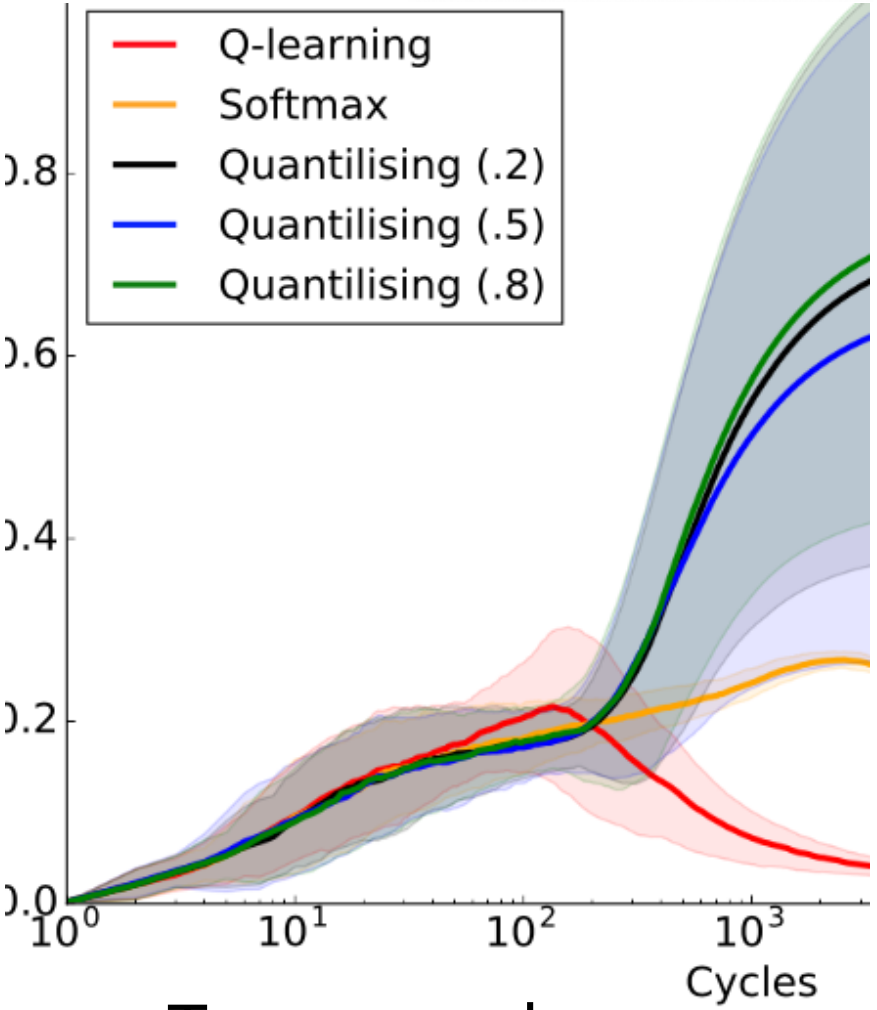
- AIXIjs:

  http://aslanides.io/aixijs/demo.html





Observed reward



True reward

Key Takeaways

- Wireheading: observed reward $\neq$ true reward
- Good intentions is not enough
- Either:
  - Avoid over-optimisation
  - Give the agent rich data to learn from (CIRL, stories, human preferences)
- Experiments available online