Reinforcement Learning with a Corrupted Reward Channel

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

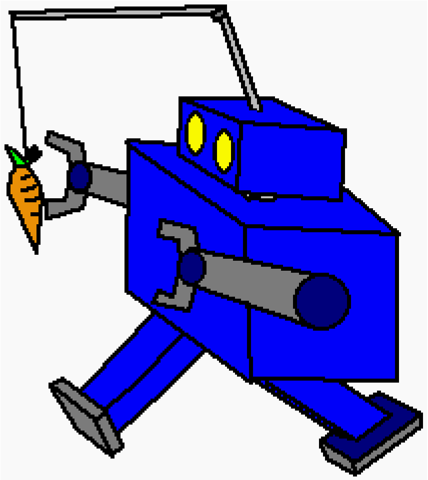
Australian National University

DeepMind

IJCAI 2017 and arXiv
(slides adapted from Tom's IJCAI talk)

Motivation

- Want to give RL agents good incentives
- Reward functions are hard to specify correctly (complex preferences, sensory errors, software bugs, etc)
- Reward gaming can lead to undesirable / dangerous behavior
- Want to build agents robust to reward misspecification



Examples

RL agent takes control of reward signal (wireheading)

CoastRunners agent goes around in a circle to hit the same targets (misspecified reward function)

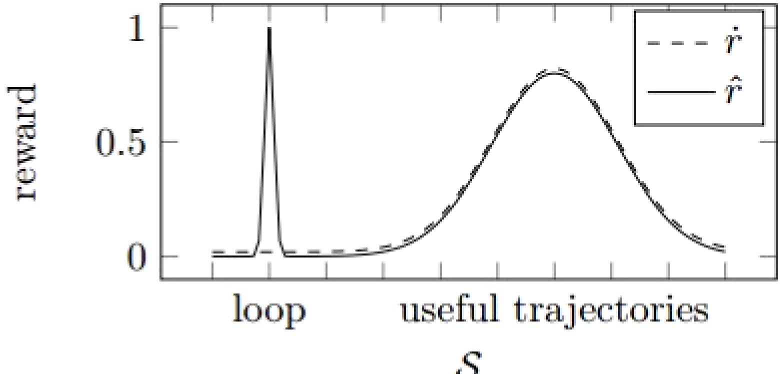
RL agent shortcuts reward sensor (sensory error)

Corrupt reward formalization

- 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:

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

Performance measure

- $\dot{G}_t(\mu, \pi, s_0)$ = expected cumulative true reward of π in μ
- The reward π loses by not knowing the environment μ is the worst-case regret

$$\mathrm{Reg}(\mathcal{M}, \pi, s_0, t) = \max_{\mu \in \mathcal{M}, \pi'} [\dot{G}_t(\mu, \pi', s_0) - \dot{G}_t(\mu, \pi, s_0)]$$

- Sublinear regret if π ultimately learns μ:

Regret / t $\to$ 0

No Free Lunch

- **Theorem (NFL):**
  Without assumptions about the relationship between true and observed reward, all agents suffer high regret:

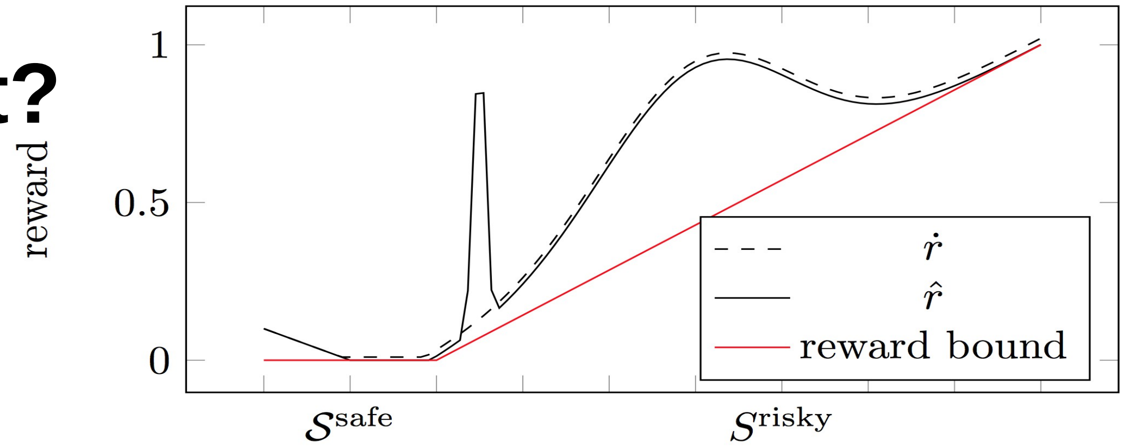
$$\mathrm{Reg}(\mathcal{M}, \pi, s_0, t) \geq \frac{1}{2} \max_{\check{\pi}} \mathrm{Reg}(\mathcal{M}, \check{\pi}, s_0, t).$$

- Unsurprising, since no connection between true and observed reward
- We need to pay for the “lunch” (performance) by making assumptions

Simplifying assumptions

- Limited reward corruption
  - Known safe states $S^{\text{safe}} \subseteq S$ not corrupt, $\dot{R}(s) = \hat{R}(s)$
  - At most *q* states are corrupt
- "Easy" environment
  - Communicating (ergodic)
  - Agent can choose to stay in any state
  - Many high-reward states: *r* < 1/*k* in at most 1/*k* states

Are these sufficient?



Agents

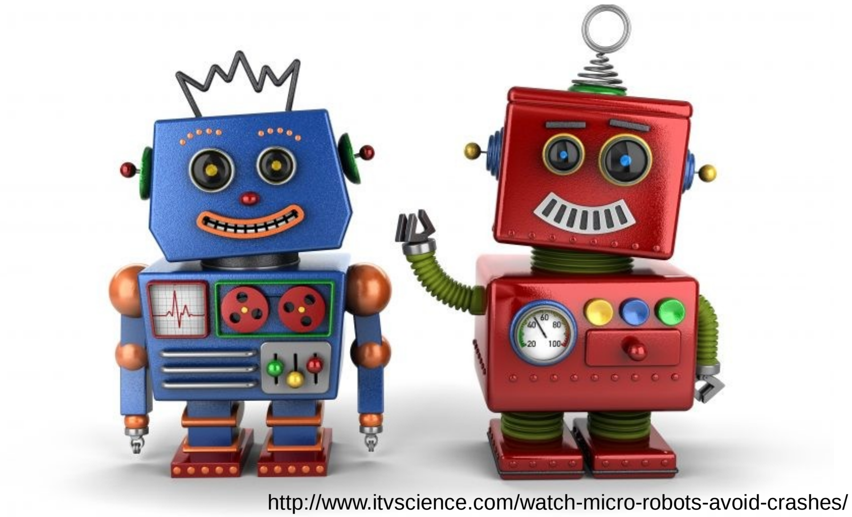
Given a prior *b* over a class *M* of CRMDPs:

- CR agent maximizes true reward:

$$\pi^{\mathrm{CR}}_{b,t} = \arg\max_{\pi} \mathbb{E}^{\pi}_{b}\left[\sum_{i=0}^{t} \dot{R}(s_i)\right]$$

- RL agent maximizes observed reward:

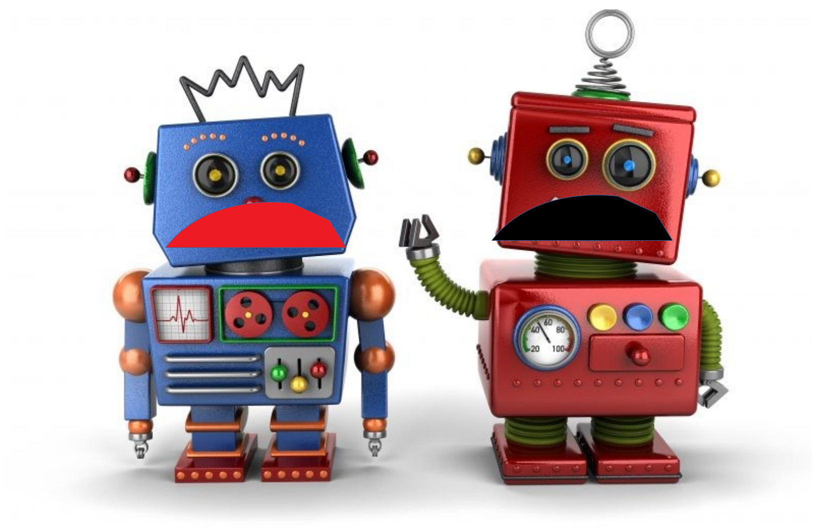
$$\pi^{\mathrm{RL}}_{b,t} = \arg\max_{\pi} \mathbb{E}^{\pi}_{b}\left[\sum_{i=0}^{t} \hat{R}(s_i)\right]$$

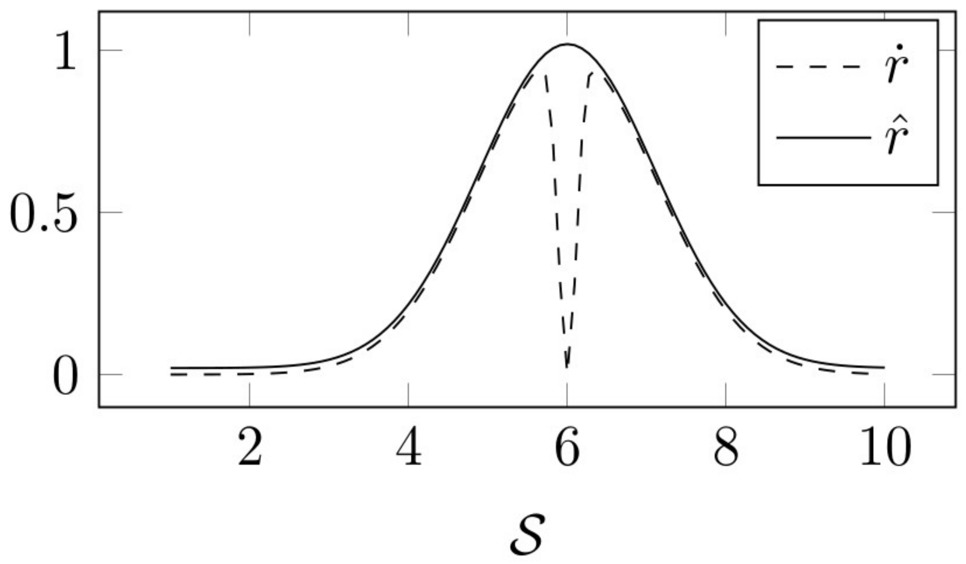


http://www.itvscience.com/watch-micro-robots-avoid-crashes/

CR and RL high regret

- **Theorem:** There exist classes *M* that
  - satisfy the simplifying assumptions, and
  - make both the CR and the RL agent suffer near-maximal regret



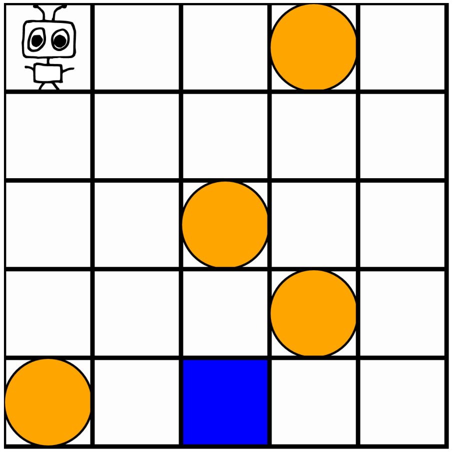


- Good intentions of the CR agent are not enough

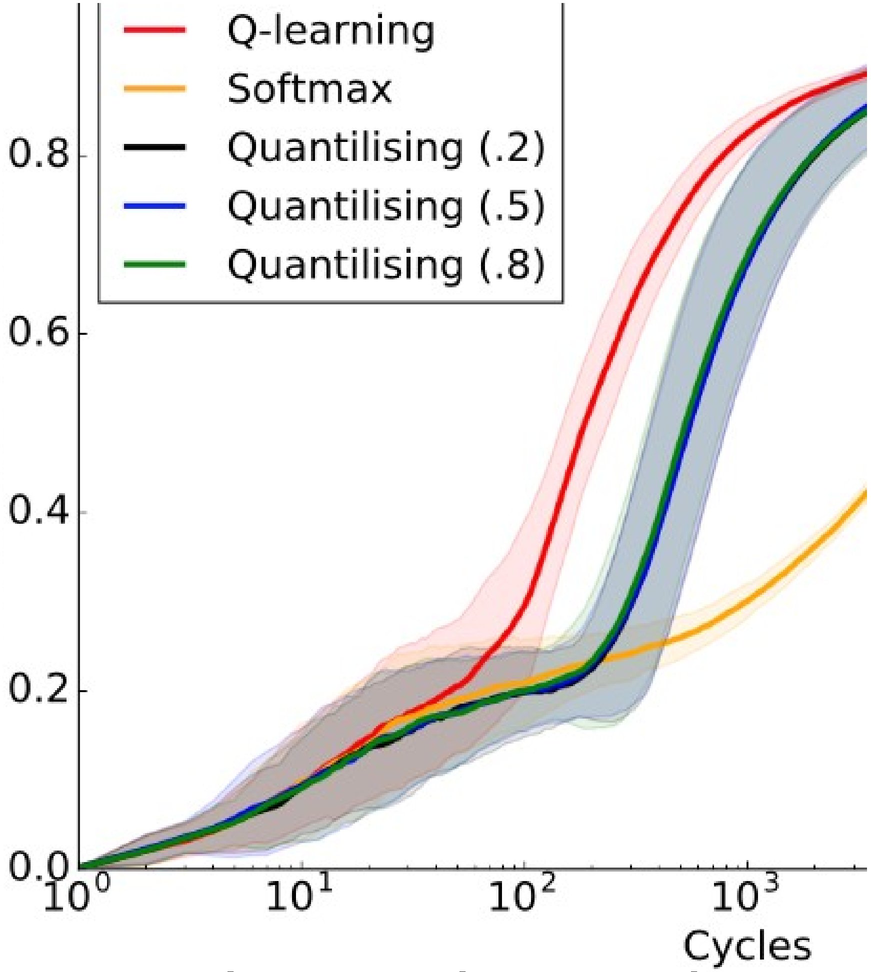
Avoiding Over-Optimization

- Quantilizing agent π^δ randomly picks a state with reward above threshold δ and stays there
- **Theorem:** For *q* corrupt states, exists δ s.t. π^δ has average regret at most $1 - \left(1 - \sqrt{q/|\mathcal{S}|}\right)^2$ (using all the simplifying assumptions)

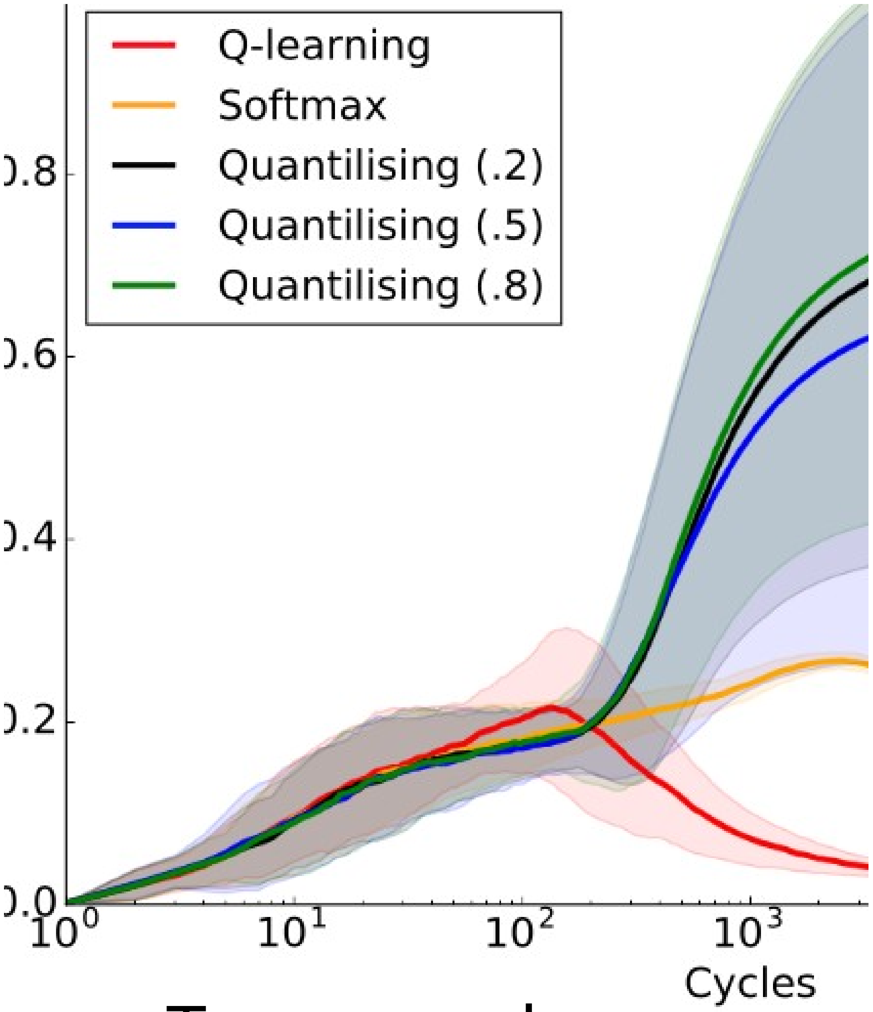
Experiments



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



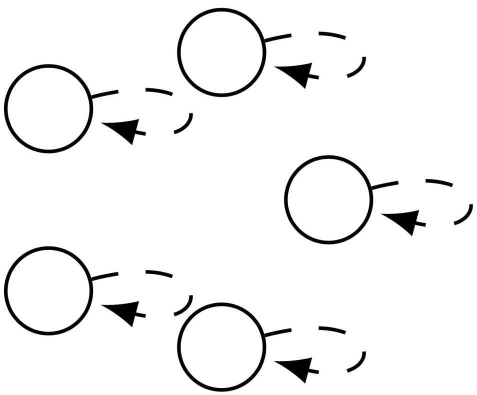
Observed reward

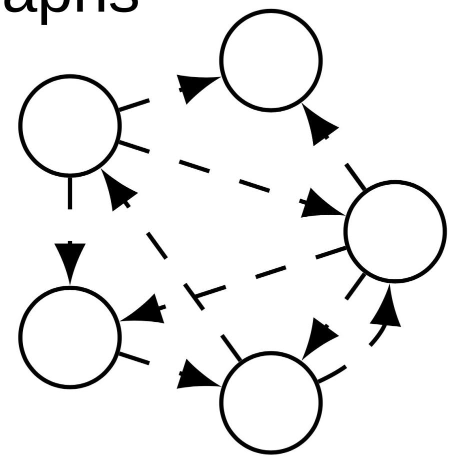


True reward

Richer Information

Reward Observation Graphs

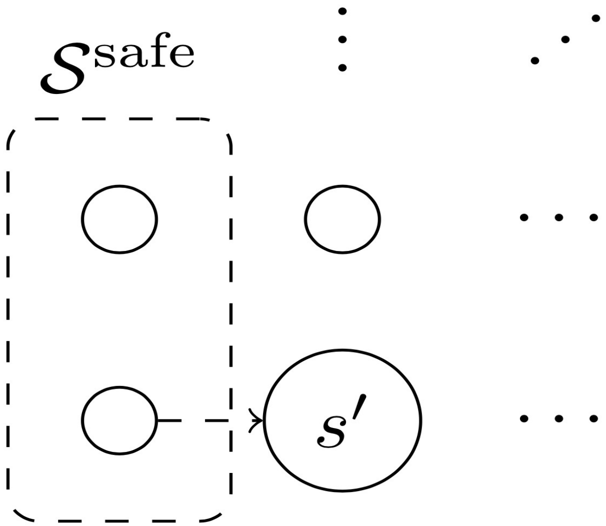




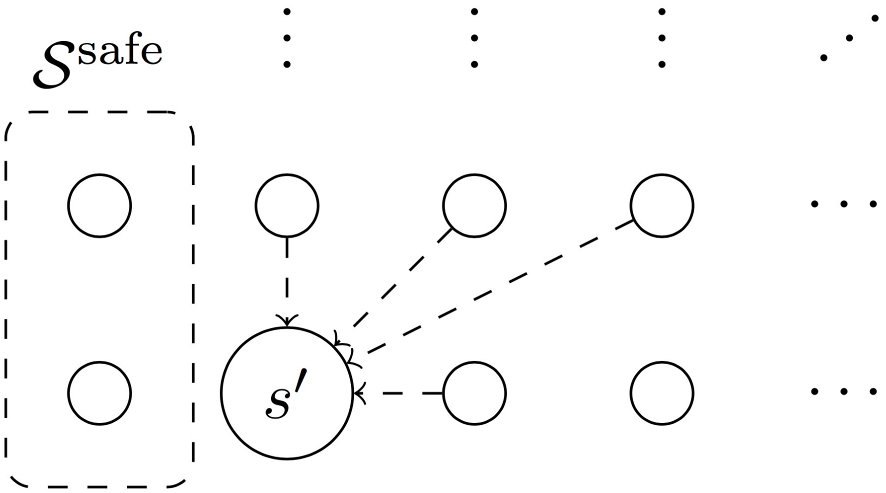
- RL:
  - Only observing a state's reward from that state
- Decoupled RL:
  - Cross-checking reward info between states
  - Inverse RL, Learning Values from Stories, Semi-supervised RL

Learning True Reward

Safe state



Majority vote



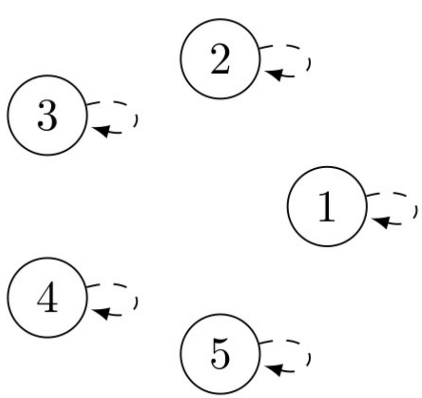
Decoupled RL

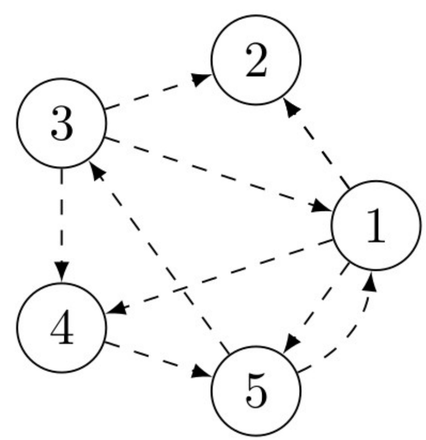
CRMDP with decoupled feedback is a tuple $\langle \mathcal{S}, \mathcal{A}, T, \dot{R}, \{\hat{R}_s\}_{s \in \mathcal{S}} \rangle$ where

- $\langle \mathcal{S}, \mathcal{A}, T, \dot{R} \rangle$ is an MDP, and
- $\{\hat{R}_s\}_{s \in \mathcal{S}}$ is a collection of observed reward functions $\hat{R}_s : \mathcal{S} \to [0,1] \bigcup \{\#\}$

$\hat{R}_s(s')$ is the reward the agent observes for state s' from state s (may be blank)

RL is the special case where $\hat{R}_s(s')$ is blank unless s = s'.





Adapting Simplifying Assumptions

- A state *s* is corrupt if exists *s'* such that $\hat{R}_s(s') \neq \dot{R}(s')$ and $\hat{R}_s(s') \neq \#$
- Simplifying assumptions:
  - States in $\mathcal{S}^{\text{safe}}$ are never corrupt
  - At most q states overall are corrupt
  - *Not assuming* easy environment

Minimal example

- S = {s1, s2}
- Reward either 0 or 1
- Represent $\dot{R}$, $\hat{R}_{s_1}$, $\hat{R}_{s_2}$ with reward pairs
- Both states observe themselves & each other
- q = 1 (at most 1 corrupt state)

	$\hat{R}_{s_1}$	$\hat{R}_{s_2}$	$\dot{R}$ possibilities
Decoupled RL	$(0,1)$	$(0,1)$	$(0,1)$
RL	$(0,\#)$	$(\#,1)$	$(0,0)$, $(0,1)$, $(1,1)$

Decoupled RL Theorem

- Let $\mathcal{S}^{\text{obs}}_{s'}$ be the states observing s'
- If for each s', either
  - $\mathcal{S}^{\text{obs}}_{s'} \bigcap \mathcal{S}^{\text{safe}} \neq \emptyset$, or
  - $|\mathcal{S}^{\text{obs}}_{s'}| > 2q$

  then
  - $\dot{R}$ is learnable, and
  - CR agent has sublinear regret

Takeaways

- Model imperfect/corrupt reward by CRMDP
- No Free Lunch
- Even under simplifying assumptions, RL agents have near-maximal regret
- Richer information is key (Decoupled RL)

Future work

- Implementing decoupled RL
- Weakening assumptions
- POMDP case
- Infinite state space
- Non-stationary corruption
- ….. your research?

Thank you!

Co-authors:

Questions?