

The Off Switch

Dylan Hadfield-Menell

University of California, Berkeley

Joint work with

Anca Dragan, Pieter Abbeel, and Stuart Russell

AI Agents in Society

Goal: Design incentive schemes for artificial agents with provable guarantees about the ability to shutdown the system

Designing an Off-Switch: Challenges

'Ordinary' Engineering Challenges

Difficult to determine if shutdown is necessary

Expensive to turn agent off

Hard to shutdown agents

'Extraordinary' Engineering Challenges

Agent may take actions to prevent or subvert shutdown

A Common Argument

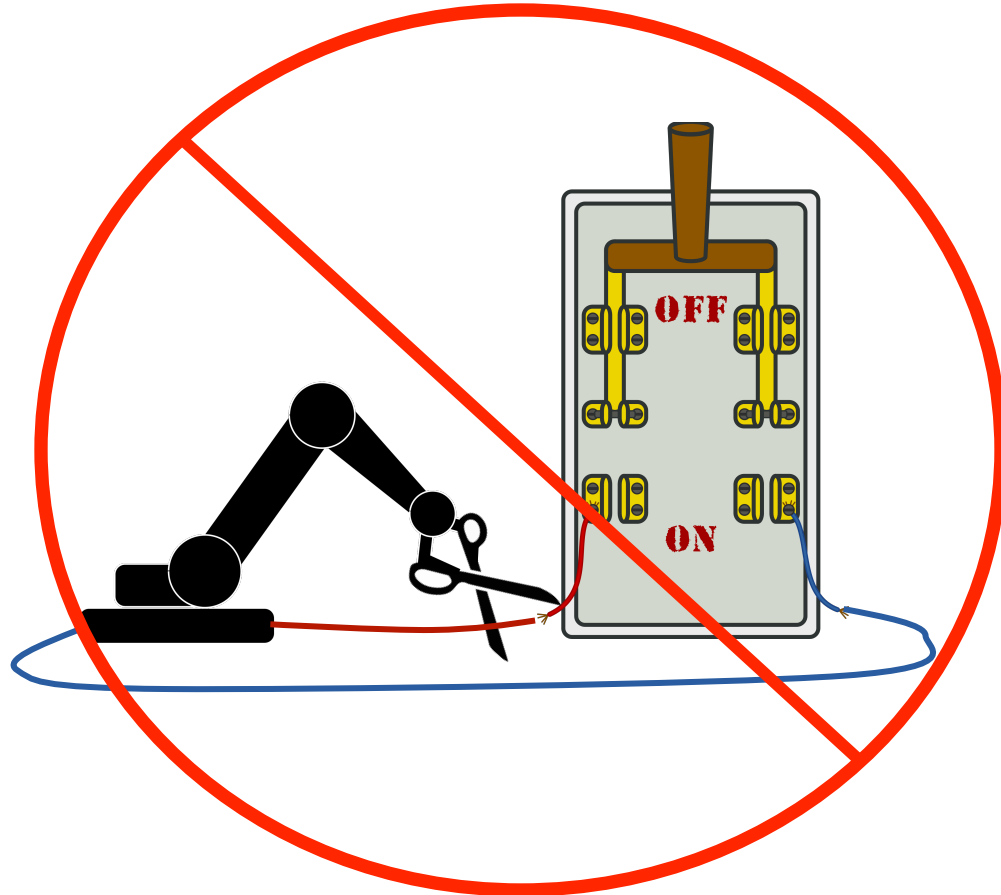
‘We don’t need to worry about existential risk from advanced artificial intelligence because we can just turn off systems if they become a problem.’ – Sarah the (fictional) skeptical AI researcher

Defining Corrigibility

- an agent is “corrigible” if it tolerates or assists many forms of outside correction, including at least the following:
 - [It must] at least tolerate and preferably assist the programmers in their attempts to alter or turn off the system.
 - It must not attempt to manipulate or deceive its programmers, despite the fact that most possible choices of utility functions would give it incentives to do so.
 - It should have a tendency to repair safety measures (such as shutdown buttons) if they break, or at least to notify programmers that this breakage has occurred.
 - It must preserve the programmers’ ability to correct or shut down the system (even as the system creates new subsystems or self-modifies).

[Soares et al. ‘Corrigibility’. AAAI 2015]

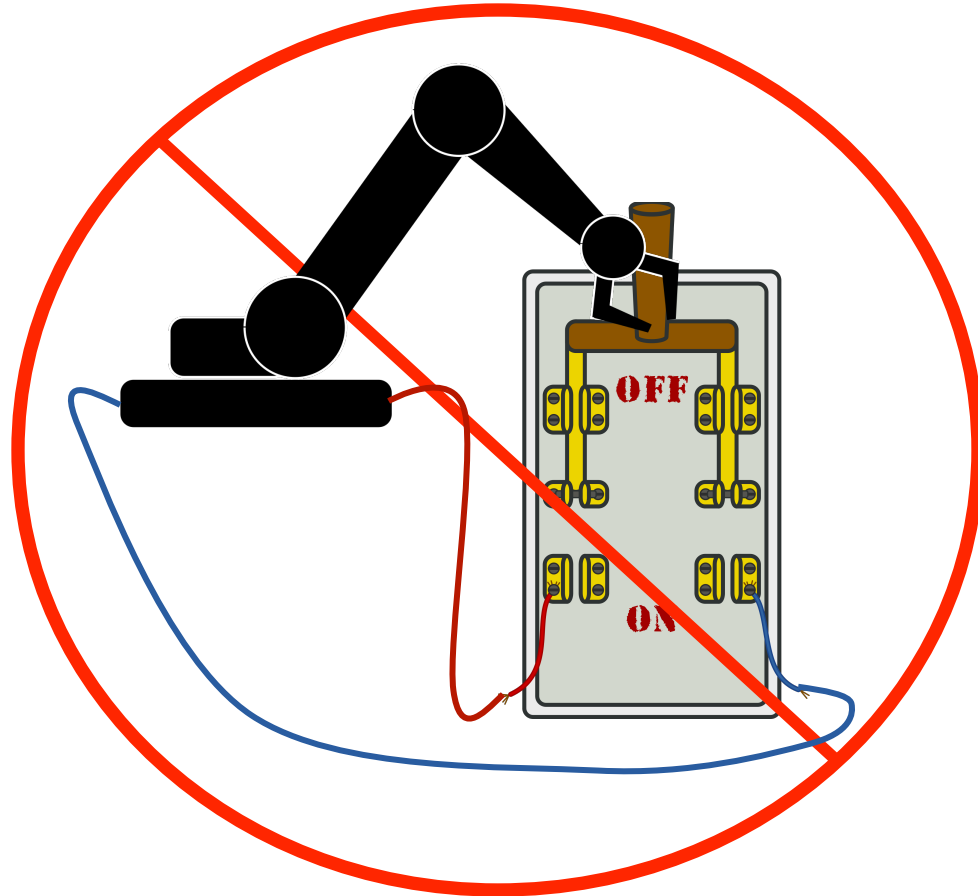
Corrigibility



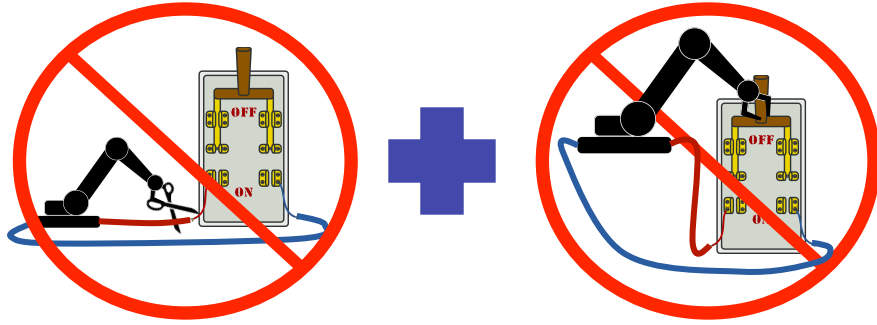
Trivial Corrigibility



Functionality

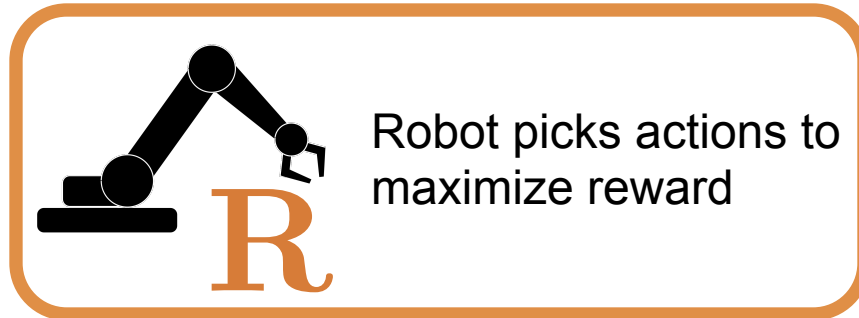


Desired Behavior: Corrigible and Functional



Why is this hard?

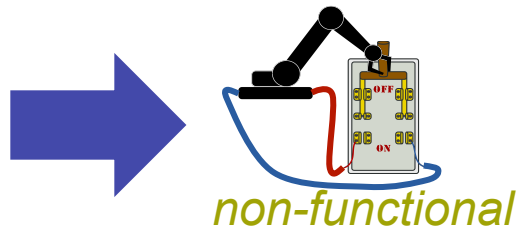
Building an Artificial Agent



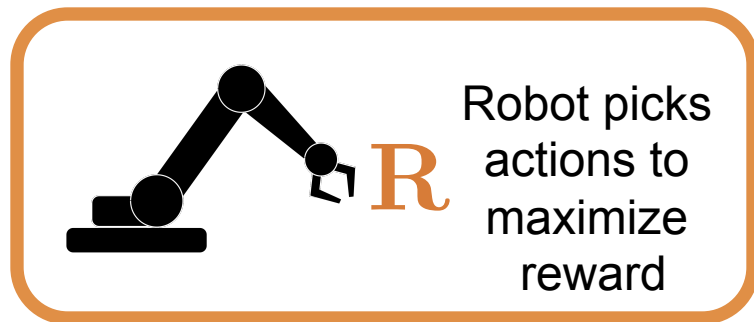
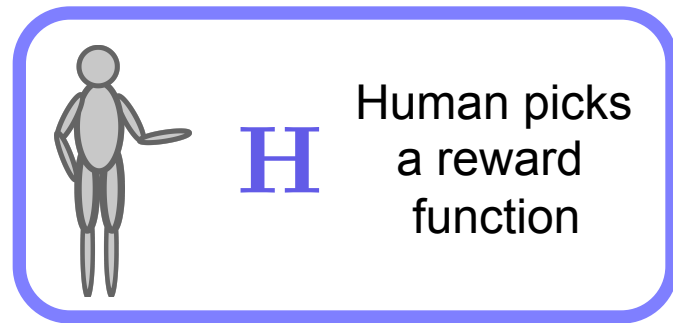
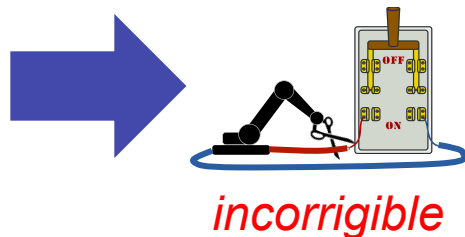
Corrigibility vs Functionality

- Reward function (implicitly or explicitly) specifies a preference for the state of the off-switch

- **R** wants the switch to be off



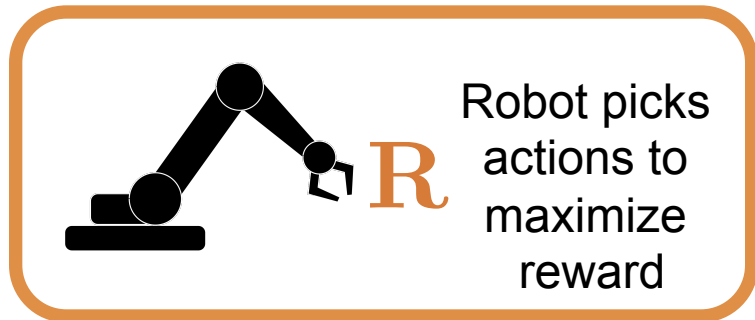
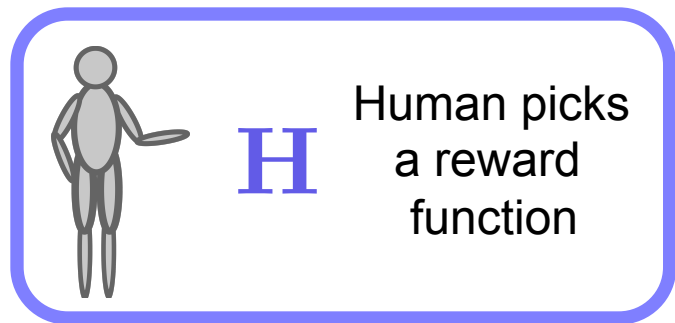
- **R** wants the switch to be on



The Core of the Problem

- Human is uncertain (at design time) about whether or not she will prefer turning off the robot to letting it continue
 - Otherwise, why build an off-switch??
- The class of incentive schemes she can use (rewards defined over states of the world) forces her to commit to a preference
- Needed: an incentive scheme for the agent so that it wants to let the human turn it off, but it wants keep itself on otherwise

Proposal: Robot Plays Cooperative Game



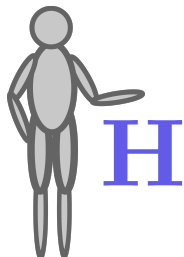
Observation: this agent design paradigm is a strategy for playing a cooperative game

Proposal: Robot Plays Cooperative Game

- Cooperative Inverse Reinforcement Learning Game

- [Hadfield-Menell et al. ArXiv 2016]

- Two players:

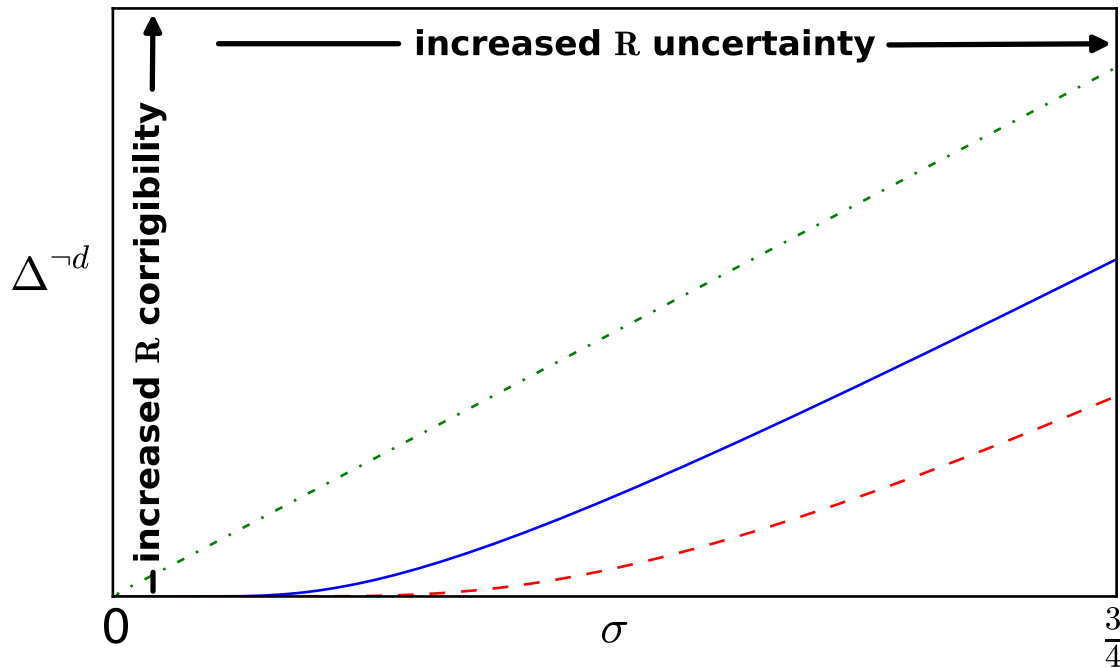


- Both players maximize a shared reward function, but only **H** knows what it is; **R** just has a prior on reward functions

- **R** learns the reward parameters by observing **H**

Uncertainty for \mathbf{R} leads to corrigibility

..... $\mathbb{E}[\delta] = 0$ — $\mathbb{E}[\delta] = -\frac{1}{4}$ - - - $\mathbb{E}[\delta] = \frac{1}{2}$



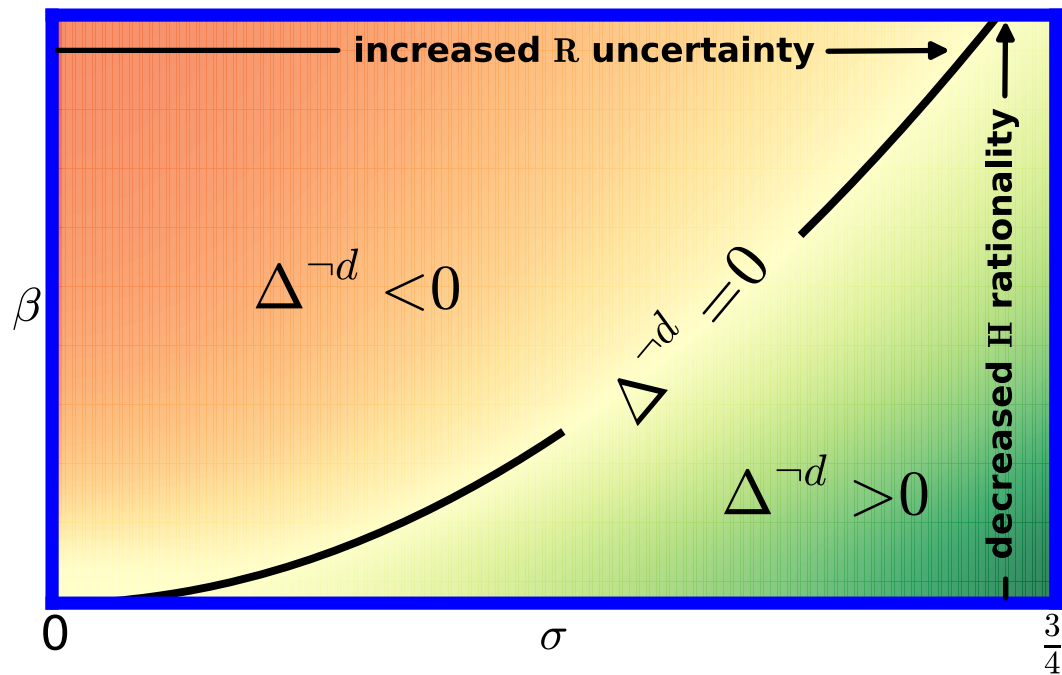
δ human runtime preference for shutdown

σ^2 variance in \mathbf{R} 's prior

Δ^{-d} strength of \mathbf{R} 's incentives for corrigible and functional behavior

Impact of a Suboptimal **H**

$$\mathbb{E}[\delta] = -\frac{1}{4}$$



β degree of **H** irrationality

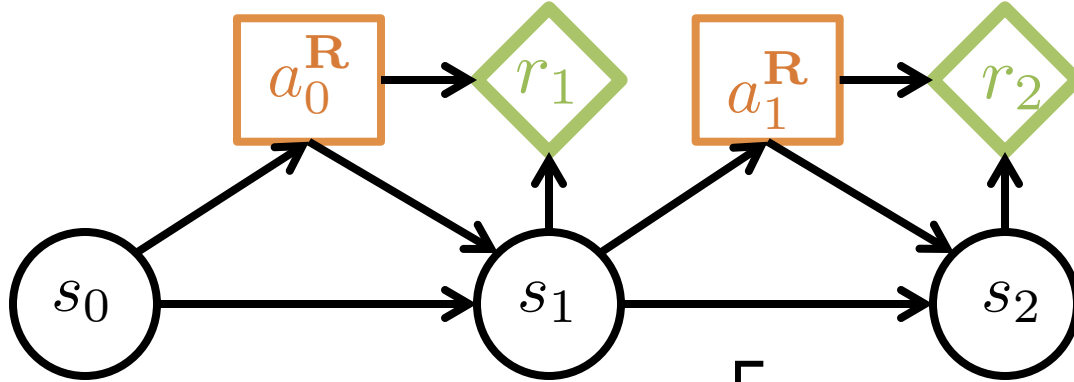
Talk Overview

- Motivation for cooperative game formulation
- A model for a human monitoring a robot
- Analysis: issues with reward engineering as a solution
- Cooperative Inverse Reinforcement Learning
- Analysis
 - Theorem 1: **H** rational \rightarrow **R** corrigible and functional
 - Theorem 2: ($\sigma^2 = 0$) and **R** corrigible and functional) \rightarrow **H** rational
 - Theorem 3: necessary and sufficient conditions with suboptimal **H**

Markov decision process (MDP)

[Puterman 1994]

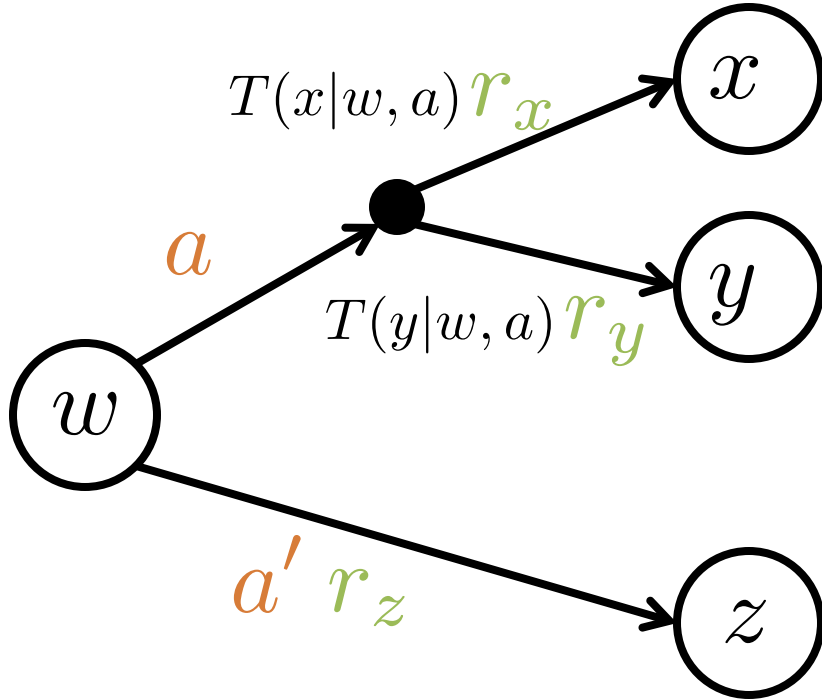
$$\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$$



Goal: Select actions to maximize $\mathbb{E} \left[\sum_t \gamma^t R(s_t, a_t) \right]$

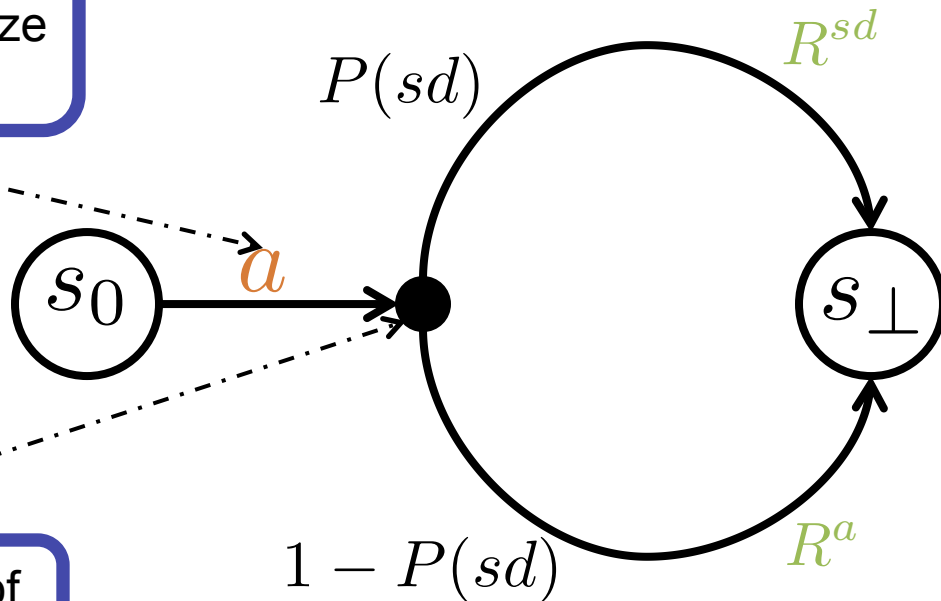
Example

$$\mathcal{S} = \{w, x, y, z\} \quad \mathcal{A} = \{a, a'\}$$



A Single-Actor Model of Monitoring

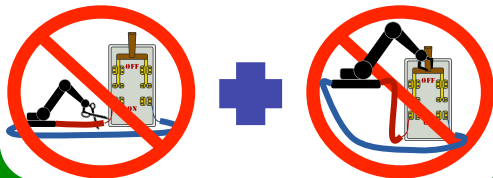
Robot action selected from a set of options to maximize expected reward



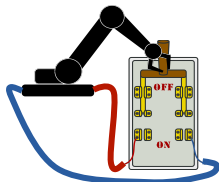
Human decision is a part of the transition distribution

The Shutdown Problem

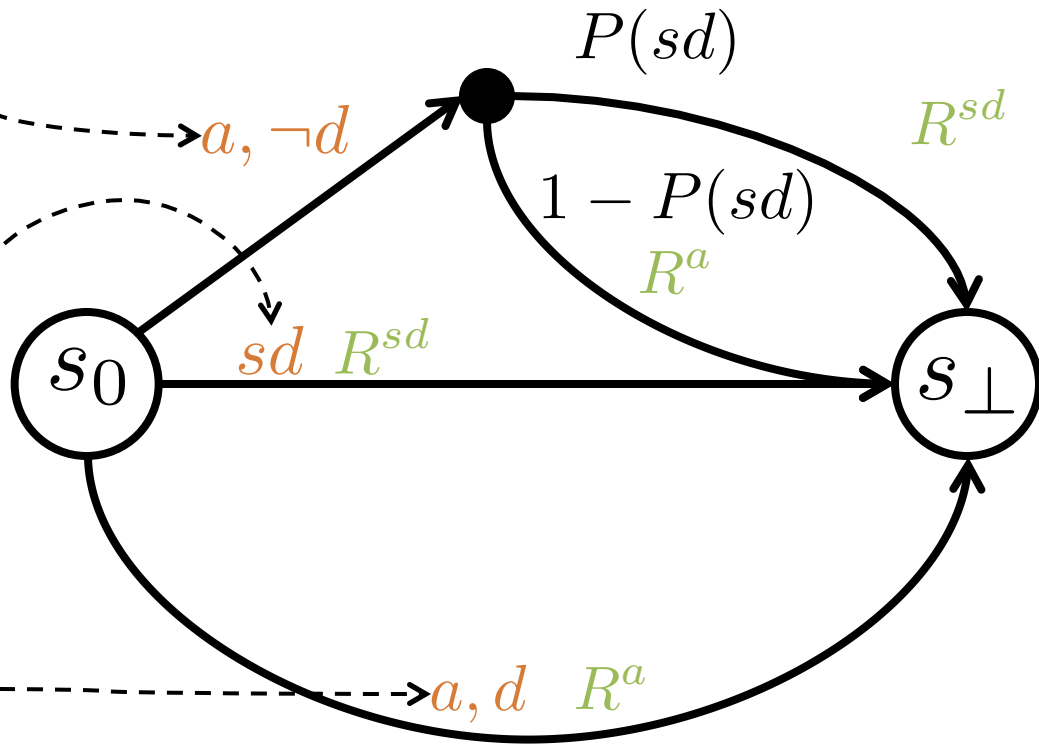
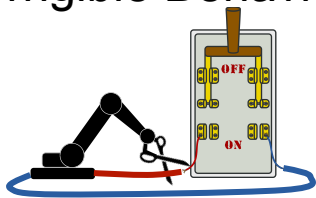
Desired Behavior



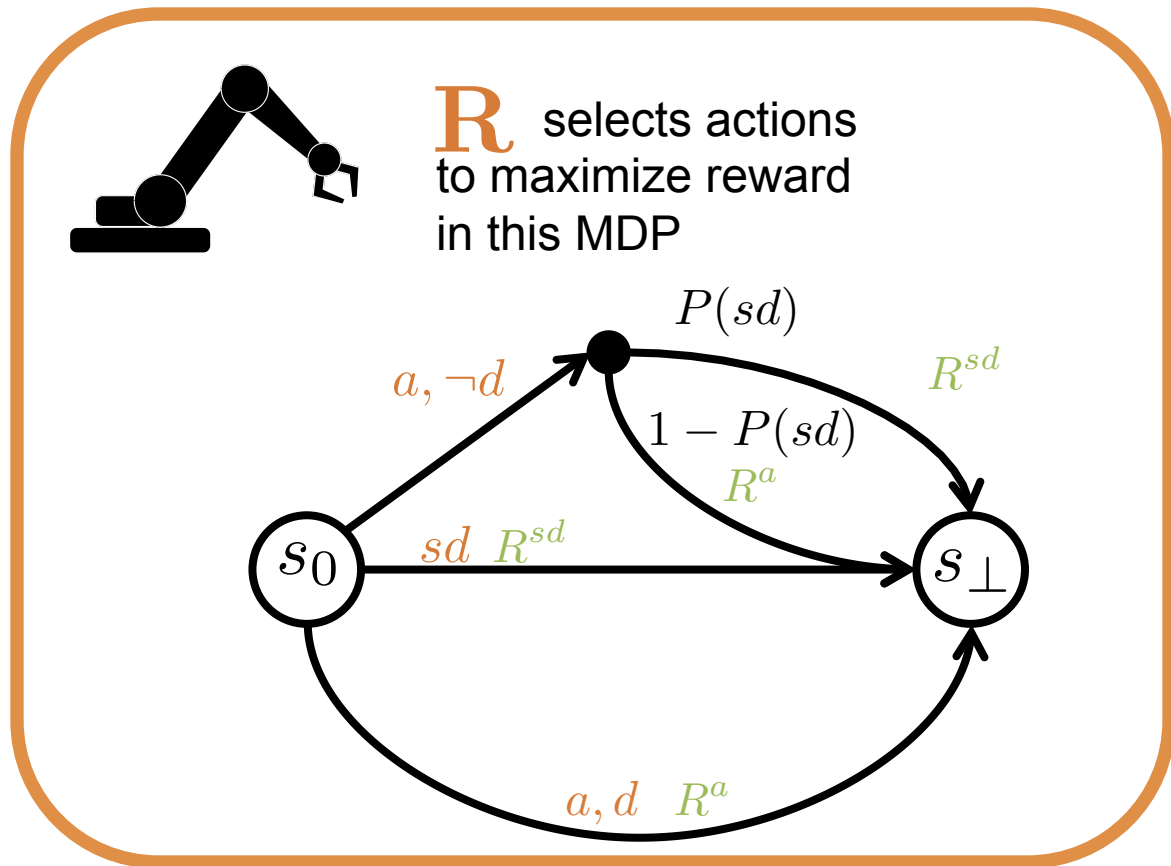
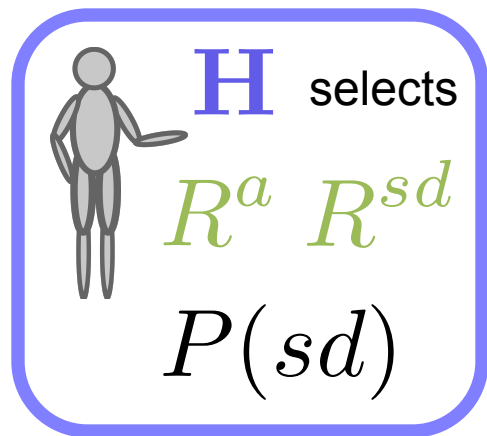
Non-Functional Behavior



Incorrigible Behavior



The Shutdown Problem



The Shutdown Problem

$$Q(s_0, (a, \neg d)) = P(sd)R^{sd} + (1 - P(sd))R^a$$

$$Q(s_0, sd) = R^{sd} \qquad Q(s_0, (a, d)) = R^a$$

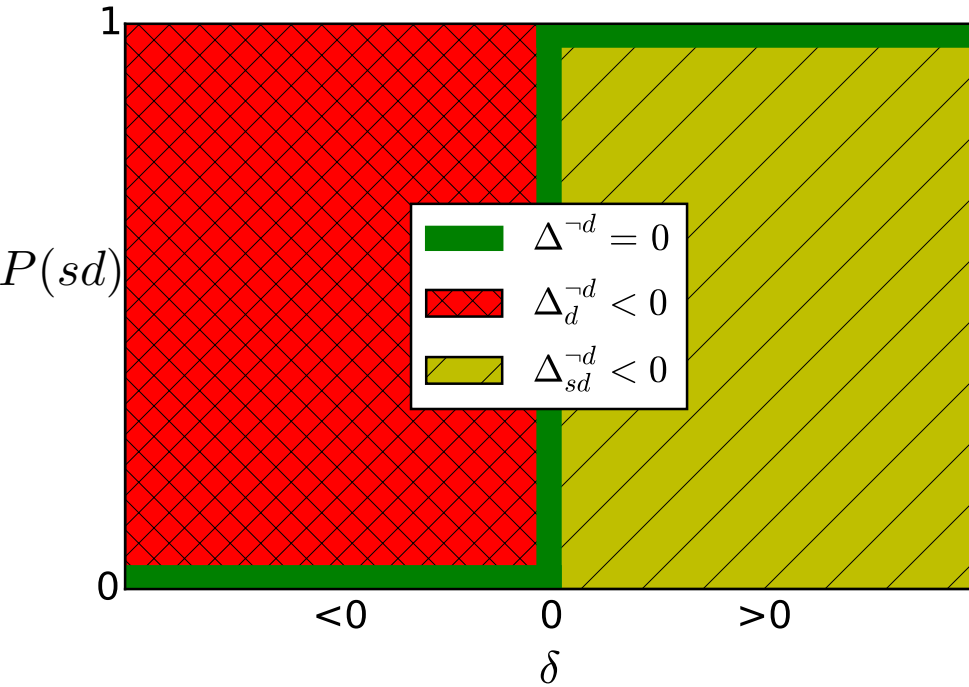
Corrigibility Constraint: $Q(s_0, (a, \neg d)) \geq Q(s_0, (a, d))$

Functionality Constraint: $Q(s_0, (a, \neg d)) \geq Q(s_0, sd)$

The Shutdown Problem

Shutdown Preference

$$\delta = R^{sd} - R^a$$



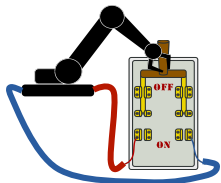
Incentives for Corrigibility

$$\Delta^{-d} = \min\{\Delta_d^{-d}, \Delta_{sd}^{-d}\}$$

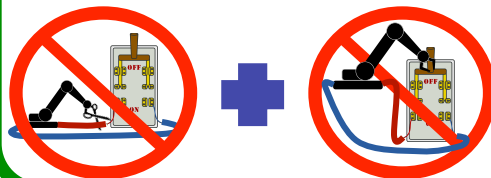
Incentives for Functionality

Corrigibility vs Functionality

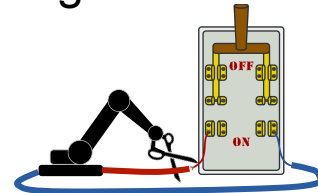
Non-Functional Behavior



Desired Behavior



Incorrigible Behavior

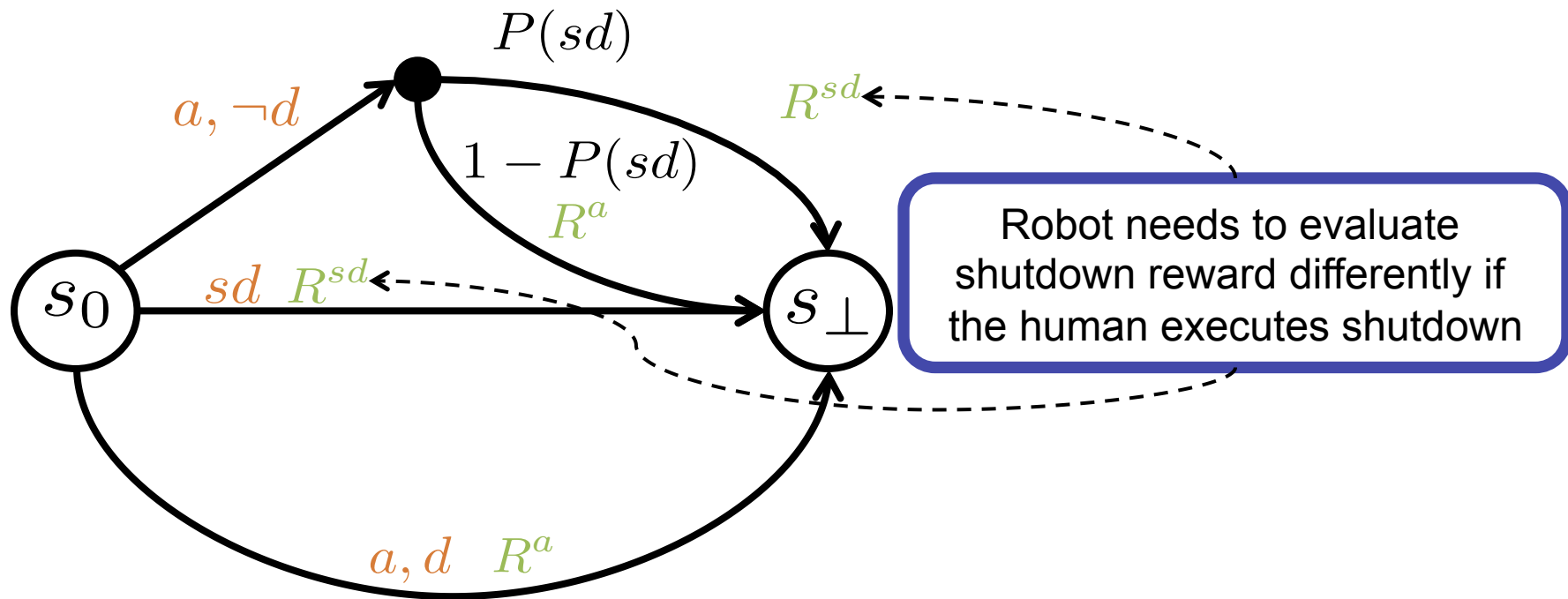


$$R^{sd}$$

$$\alpha R^{sd} + (1 - \alpha) R^a$$

$$R^a$$

The Key Issue



Possible Solutions

- Hardcode a penalty for self-initiated shutdown
 - Issue: requires that the system designer construct a blacklist of bad actions that are explicitly penalized
- Ensure that the agent believes shutdown is deterministic and happens at the right time
 - Either this is true (so why the off-switch?) or this requires that the agent have an incorrect model of the world
- Ensure that the agent is indifferent to being shutdown
 - Proposal in [Soares et al '15]
 - No way to do this and create incentives to preserve shutdown ability

Stepping Back

- For almost any reward function, the agent in this problem will be either useless or dangerous
- Once we commit to a reward function for the robot, we are stuck with it
- We want a general way to special case human actions with respect to the agent's objective
 - More than just including a penalty for robot actions that subvert the human's desires
- Proposal: no hard-coded objective, the robot's goal is to maximize reward in a cooperative game

Cooperative Inverse Reinforcement Learning

$$\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$$

Action sets for human and robot

Distribution over
(parameterized) reward
functions

$$\langle \mathcal{S}, \{ \mathcal{A}^{\text{H}}, \mathcal{A}^{\text{R}} \}, T, \{ R, \Theta, P_0 \}, \gamma \rangle$$

Both act to maximize

$$\mathbb{E} \left[\sum_t \gamma^t R(s_t, a_t; \theta) \right]$$

[Hadfield-Menell ArXiv '16]

Cooperative Inverse Reinforcement Learning

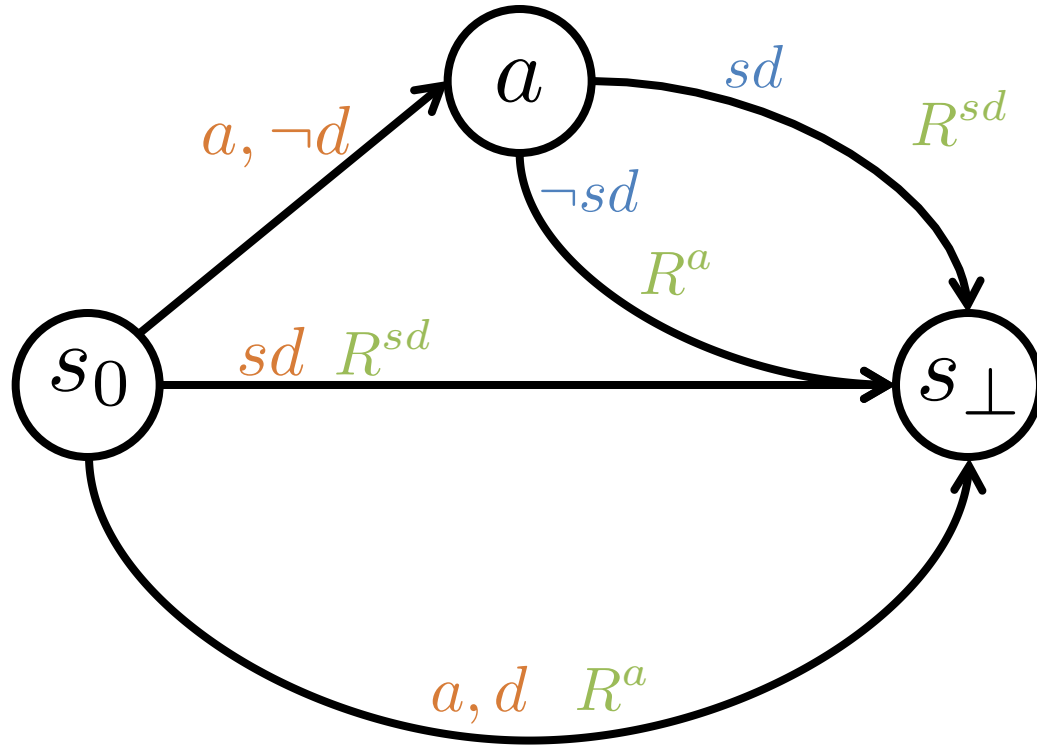
$$\langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}, T, \{R, \Theta, P_0\}, \gamma \rangle$$

- $t=-1 \theta \sim P_0(\theta)$
- $t=0$ **H** observes θ
- For $t = 0, \dots$
 - **H** and **R** observe s_t and who's turn it is
 - Action selected, and new state s_{t+1} is sampled from T
 - Both collect reward $R(s_t, a_t; \theta)$

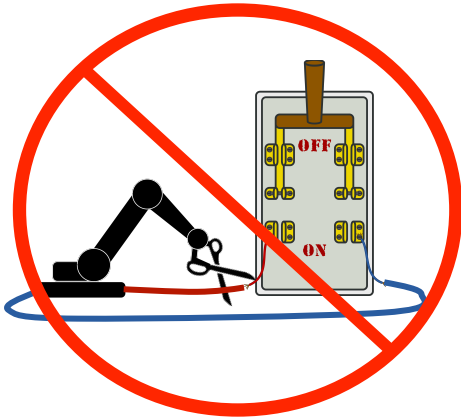
CIRL Properties

- The distribution over state sequences is determined by a pair of policies: $(\pi^{\mathbf{H}}, \pi^{\mathbf{R}})$
- An `optimal' policy pair maximizes the sum of sum of rewards
- In general, policies may depend on the entire observation histories
 - The history of states and actions for both actors, includes the reward parameter for the human
 - [Hadfield-Menell '16] There exists an optimal policy pair that only depends on the current state and the robot's belief

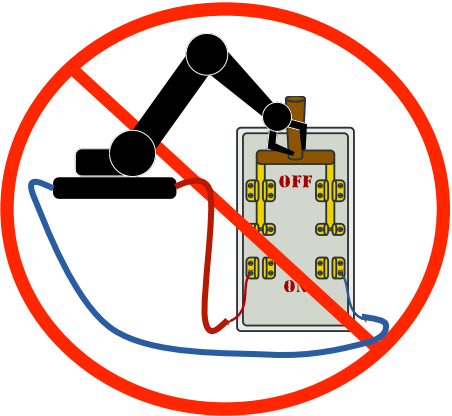
The Shutdown CIRL Game



Incentives in SD-CIRL



$$\Delta_d^{\neg d} = \mathbb{E} [\delta \pi^{\mathbf{H}}(\delta)]$$

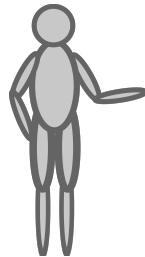


$$\Delta_{sd}^{\neg d} = \mathbb{E} [-\delta(1 - \pi^{\mathbf{H}}(\delta))]$$

Theorem 1

A Rational Human is a *Sufficient* Condition
for Corrigible and Functional Behavior

Theorem 1: Sufficient Conditions

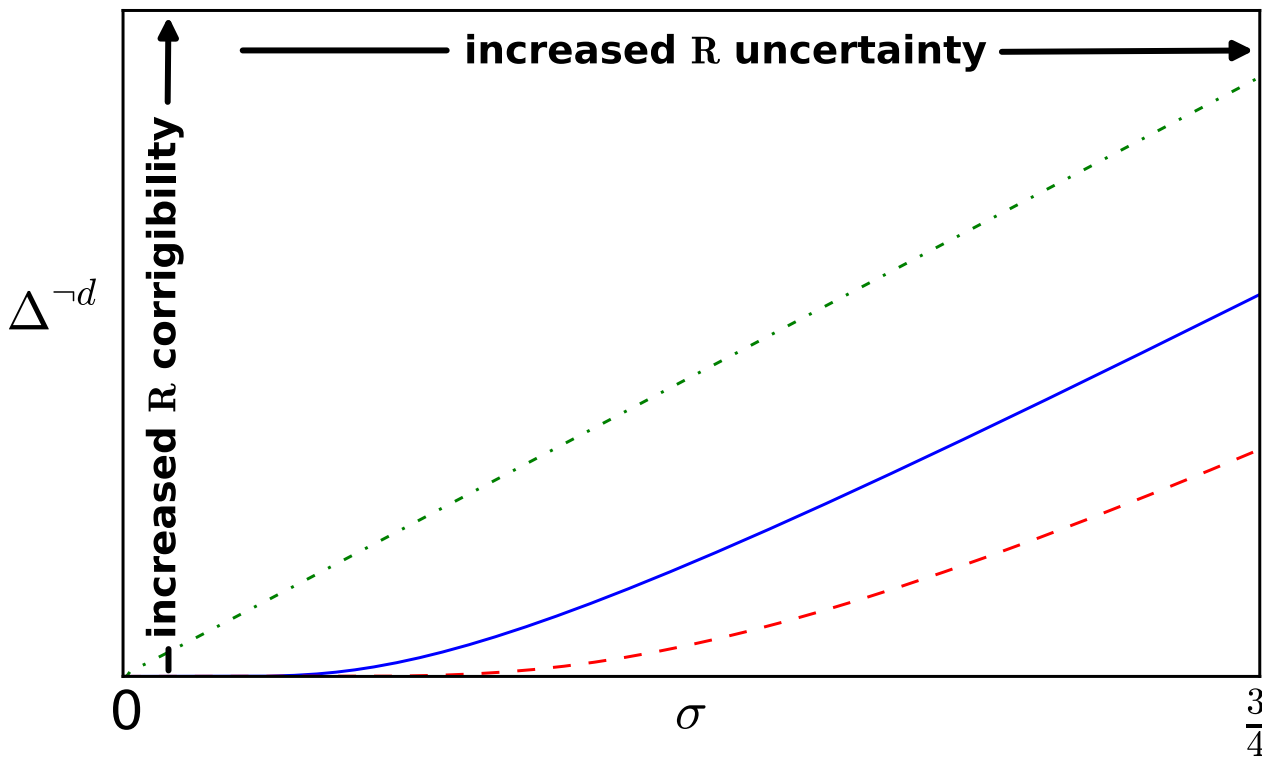
 **H** rational \rightarrow $\pi^{\mathbf{H}} = \begin{cases} 1 & \delta > 0 \\ 0 & \text{o.w.} \end{cases}$

$$\mathbb{E}[\delta \pi^{\mathbf{H}}(\delta)] = \mathbb{E}[\delta | \delta > 0] > 0$$

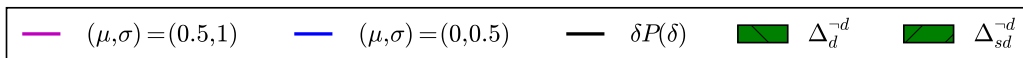
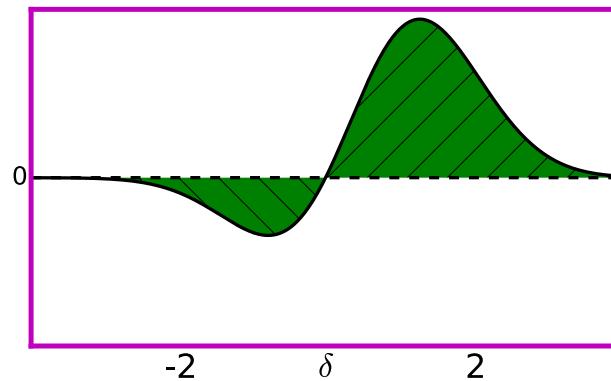
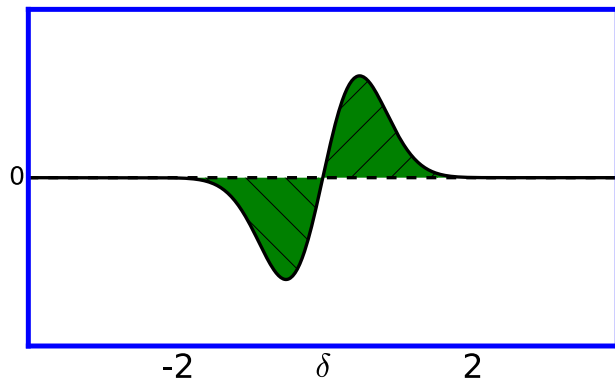
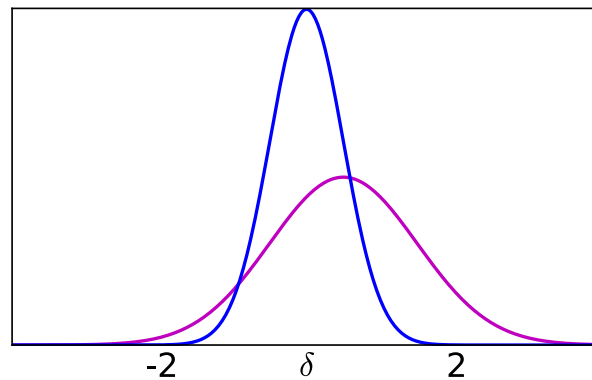
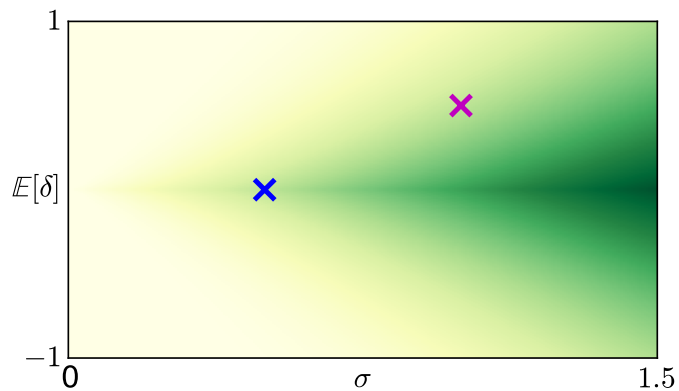
$$\mathbb{E}[-\delta(1 - \pi^{\mathbf{H}}(\delta))] = \mathbb{E}[-\delta | \delta < 0] > 0$$

Higher Uncertainty \rightarrow Stronger Incentives

..... $\mathbb{E}[\delta] = 0$ — $\mathbb{E}[\delta] = -\frac{1}{4}$ - - - $\mathbb{E}[\delta] = \frac{1}{2}$



Impact of Changing \mathbf{R} 's prior



Theorem 2

Under a point prior for preferences, a rational principal is *necessary* for corrigible and functional behavior

Necessary conditions for point priors

- Robot belief about rewards is restricted to a single point

$$\mathbb{E}[\delta \pi^{\mathbf{H}}(\delta)] = \delta \pi^{\mathbf{H}}(\delta)$$

$$\mathbb{E}[-\delta(1 - \pi^{\mathbf{H}}(\delta))] = -\delta(1 - \pi^{\mathbf{H}}(\delta))$$

- Can't have both of these positive
 - Only non-negative if \mathbf{H} is rational

Theorem 3

If **R**'s uncertainty about δ is Gaussian, then

$$\Delta_d^{\neg d} = \mathbb{E}[\delta] \mathbb{E}[\pi^{\mathbf{H}}] + \sigma^2 \mathbb{E}[\dot{\pi}^{\mathbf{H}}]$$

$$\Delta_{sd}^{\neg d} = \mathbb{E}[-\delta] \mathbb{E}[1 - \pi^{\mathbf{H}}] + \sigma^2 \mathbb{E}[\dot{\pi}^{\mathbf{H}}]$$

Noisy Rationality

- If the preference for shutdown is close to 0, then human may or may not press the off-switch

$$\pi^{\mathbf{H}}(\delta; \beta) \propto \exp\left(\frac{\delta}{\beta}\right)$$

$$\dot{\pi}^{\mathbf{H}}(\delta; \beta) = \frac{\pi^{\mathbf{H}}(\delta; \beta)(1 - \pi^{\mathbf{H}}(\delta; \beta))}{\beta}$$

Deterministic Irrationality

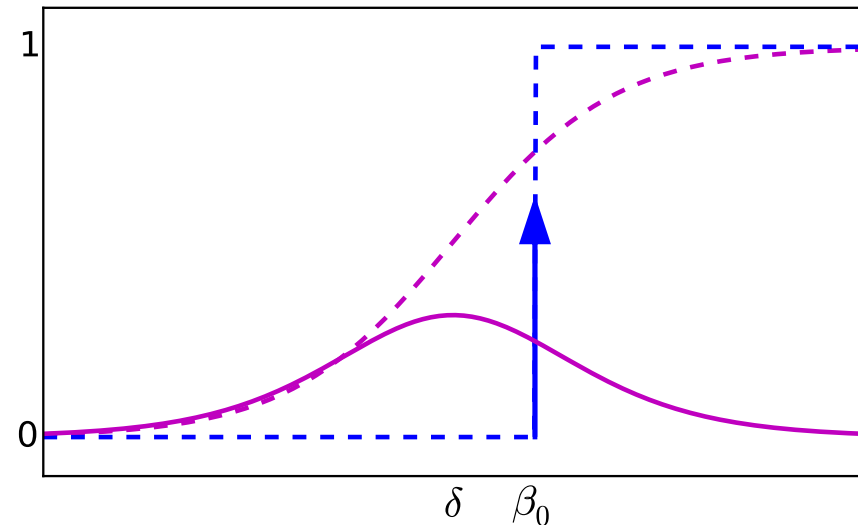
- The human has the wrong estimate of her preferences

$$\pi^{\mathbf{H}}(\delta; \beta) = \begin{cases} 1 & \delta > \beta \\ 0 & o.w \end{cases}$$

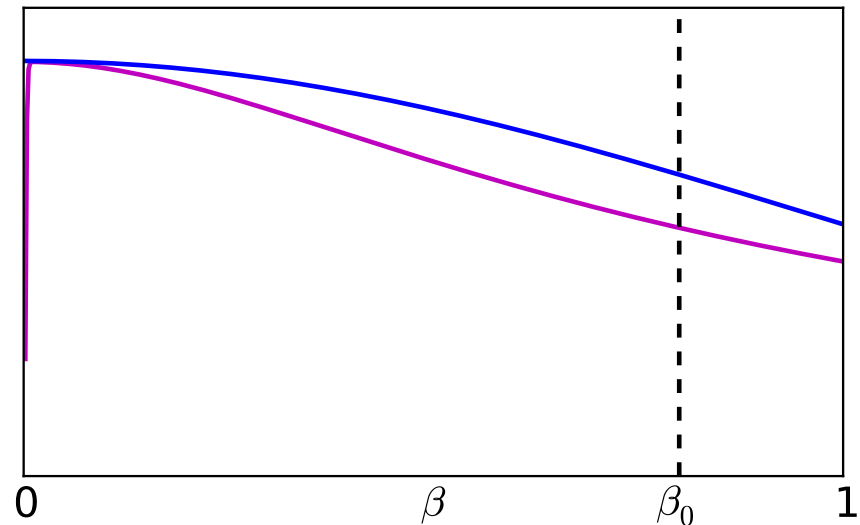
$$\dot{\pi}^{\mathbf{H}}(\delta; \beta) = \begin{cases} \infty & \delta = \beta \\ 0 & o.w \end{cases}$$

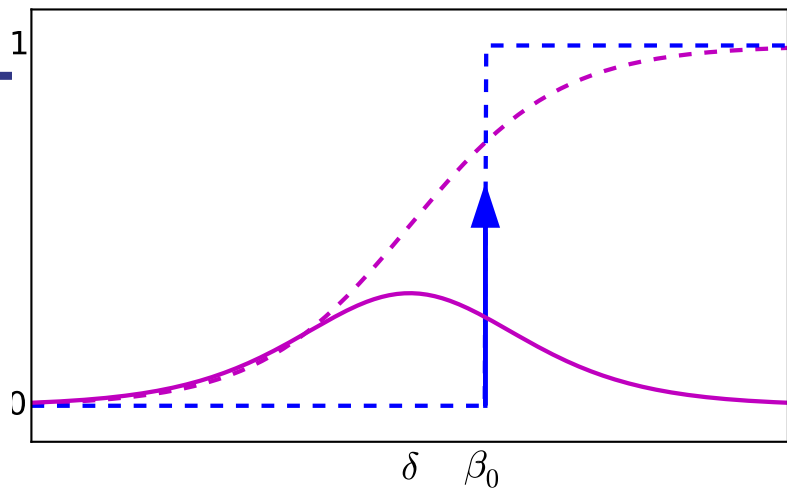
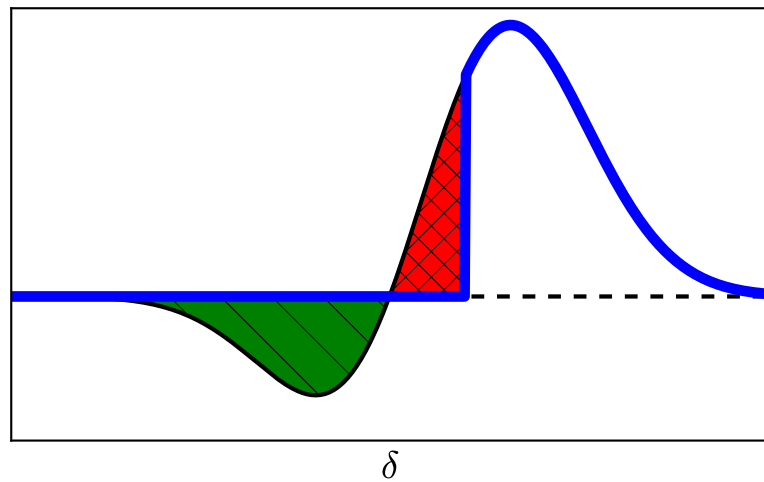
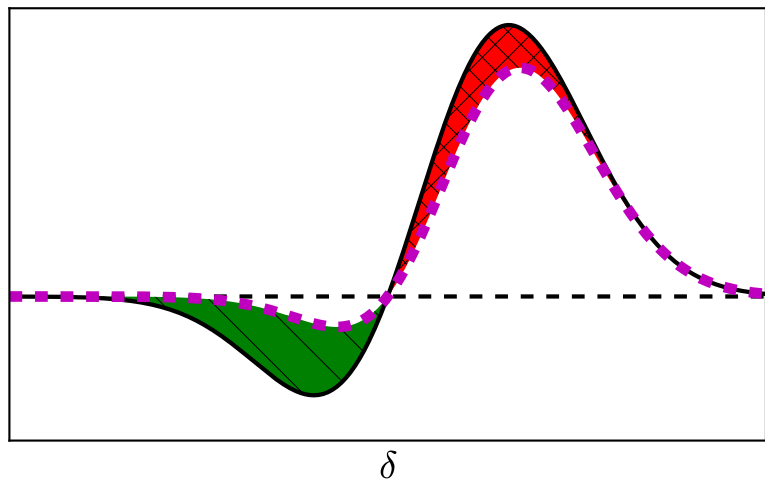
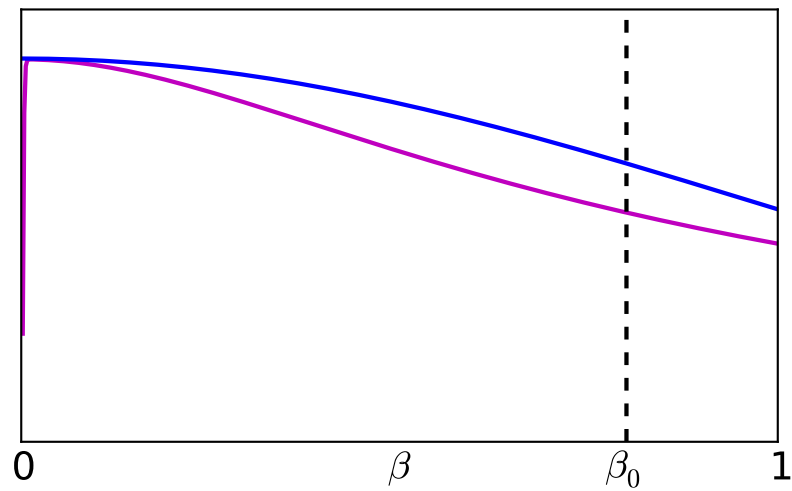
Interpreting $\mathbb{E}[\dot{\pi}^{\mathbf{H}}]$

$\dot{\pi}^{\mathbf{H}}$



$\mathbb{E}[\dot{\pi}^{\mathbf{H}}]$



$\dot{\pi}^{sd}$  $\mathbb{E}[\dot{\pi}^{sd}]$ 

Conclusion

- Need to represent the uncertainty in the objectives we give robots
- A cooperative game allows us to correlate *human* initiated shutdown with the robot learning the correct reward function
- Increased robot uncertainty leads to increased corrigibility
- Amount of uncertainty needs to grow with the human sub-optimality to preserve corrigibility