The Off Switch Game

AI Safety Reading Group @ RL Lab
Summary Slides

AI Agents in Society

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

Slide: Dylan Hadfield-Menell, 2016

Designing an Off-Switch: Challenges

‘Ordinary’ Engineering Challenges

Difficult to determine if shutdown is necessary

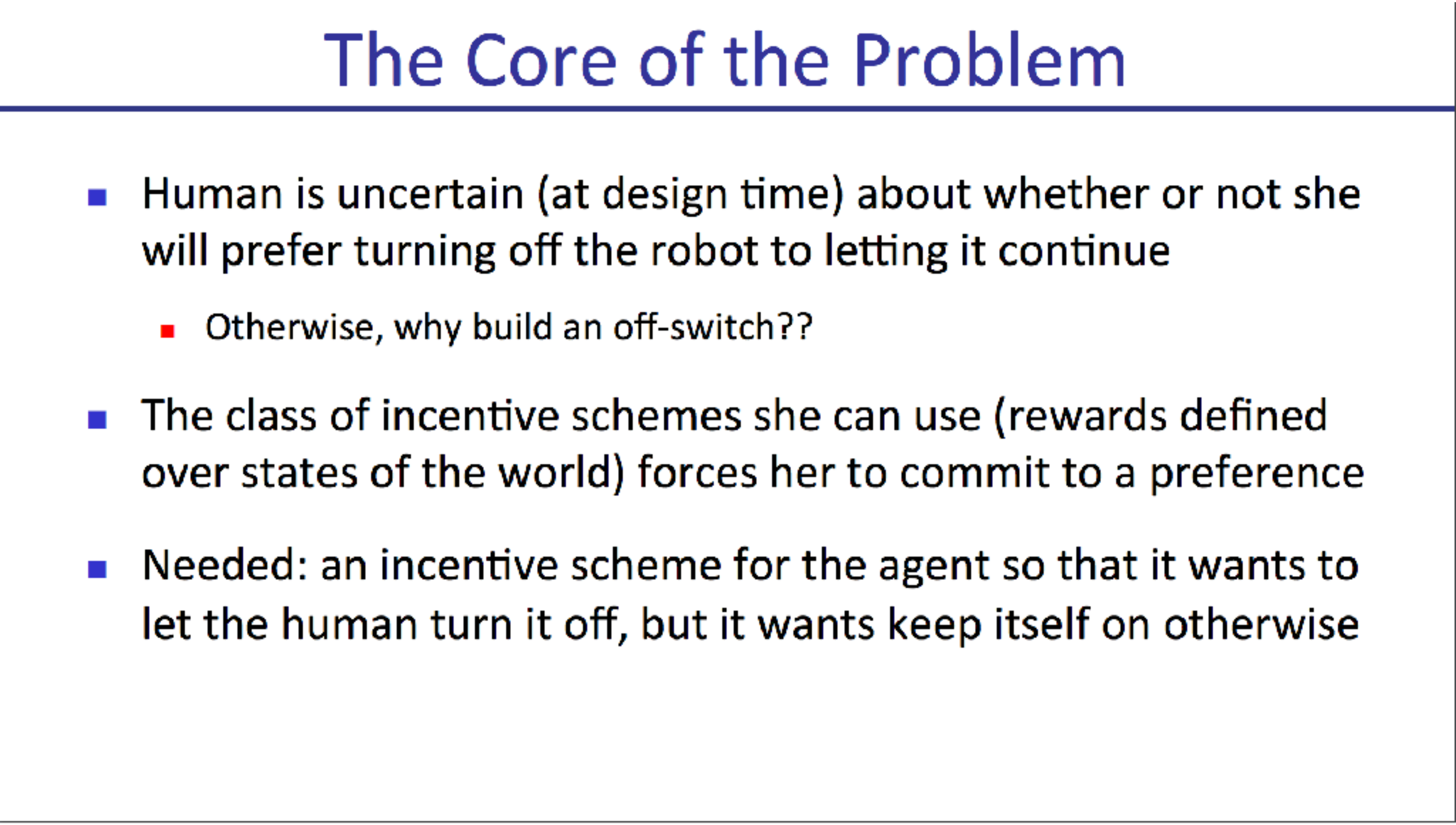
Expensive to turn agent off

Hard to shutdown agents

‘Extraordinary’ Engineering Challenges

Agent may take actions to prevent or subvert shutdown

Slide: Dylan Hadfield-Menell, 2016

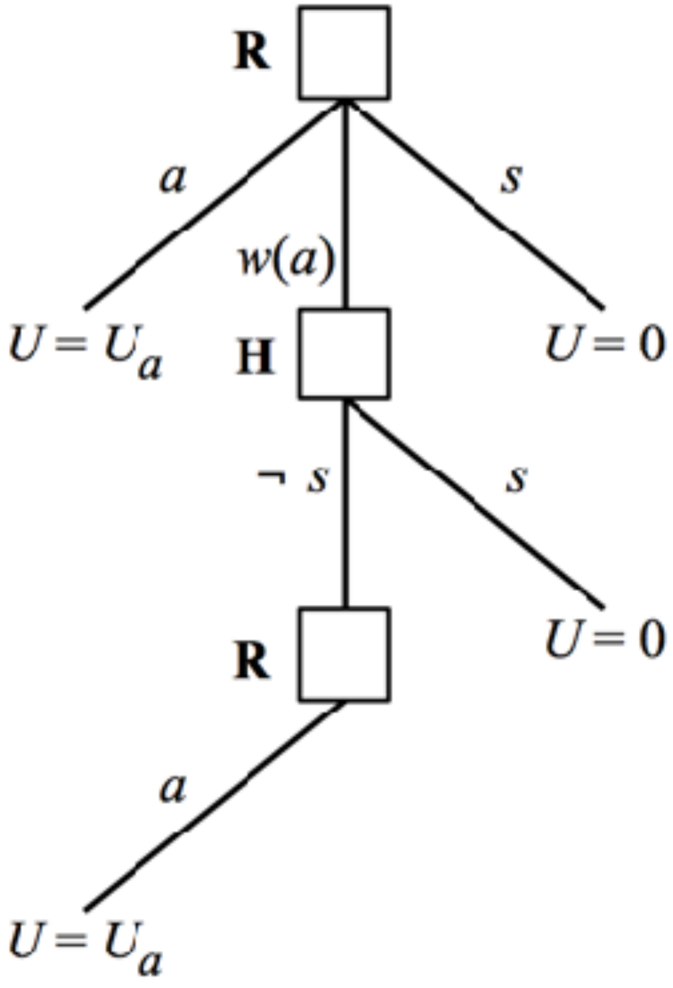


The Core of the Problem

- 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

Slide: Dylan Hadfield-Menell, 2016

Structure of the Off-Switch Game



R’s actions would fall into 1 out of 3 categories:

- Take action a, achieving utility U = Ua for H
- Take action w(a) informs H that R would like to do a and wait for H’s response
- Action s switches R off; U = 0

Formal Model

- Type of H given by Ua, assumed to be sampled by a prior $P(U_a)$

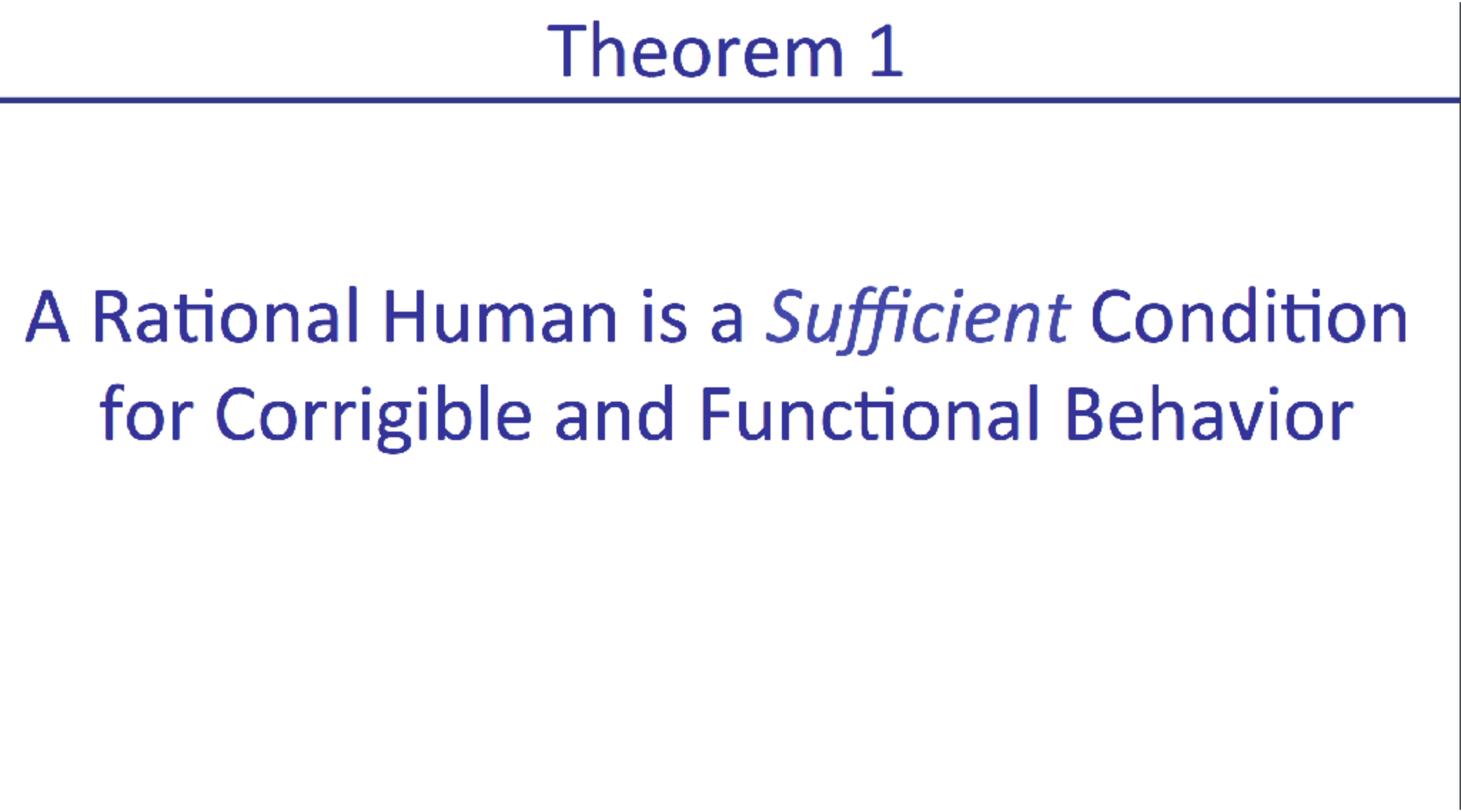
- Initial belief of R about Ua: $B^R = P(U_a)$

- ***Goal:*** *Examine the incentive that a rational R has to hold off on executing a directly and allow H to press the off switch*

- Incentive is represented as the difference in the value between a and the next best option

Incentive depends on R's belief state B^R and H's policy

$$\begin{aligned}\Delta &= \mathbb{E}[\pi^{\mathbf{H}}(U_a)U_a] - \max\{\mathbb{E}[U_a], 0\} \\ &= \min\{\mathbb{E}[-U_a(1-\pi^{\mathbf{H}}(U_a))], \quad \mathbb{E}[U_a\pi^{\mathbf{H}}(U_a)]\} \quad (1)\end{aligned}$$



Theorem 1

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

Slide: Dylan Hadfield-Menell, 2016

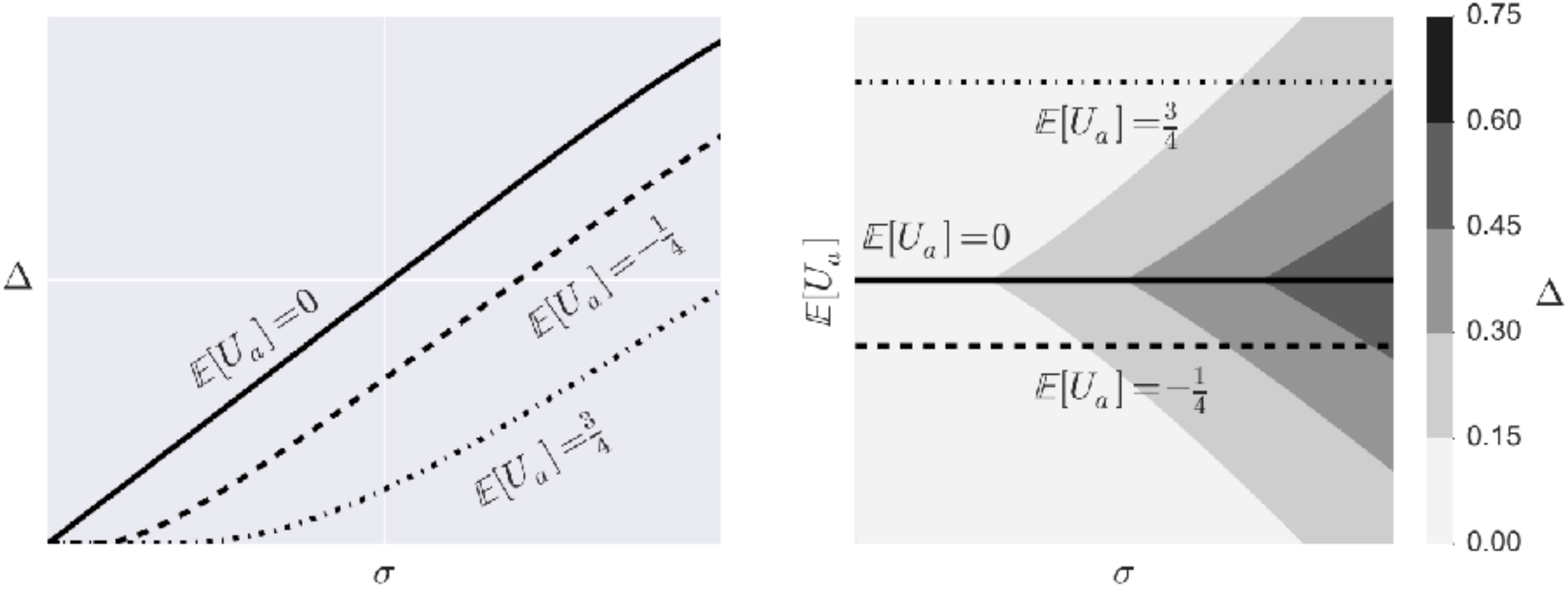
Theorem 1

If H follows a rational policy, R will allow itself to be switched off and the following holds:

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

- R's incentive to allow itself to be switched off is non-negative
- If B^R has non-zero support on the events $U_a > 0$ and $U_a < 0$, then R has a strictly positive incentive to allow itself to be switched off

Higher uncertainty -> Stronger Incentives



- Increased variance in B^R leads to an increase in incentive : Higher the uncertainty -> Stronger the incentives
- As the expected utility of a increases, R’s incentive to show a to H decreases

Suboptimal Human Decisions ?

Scenario:

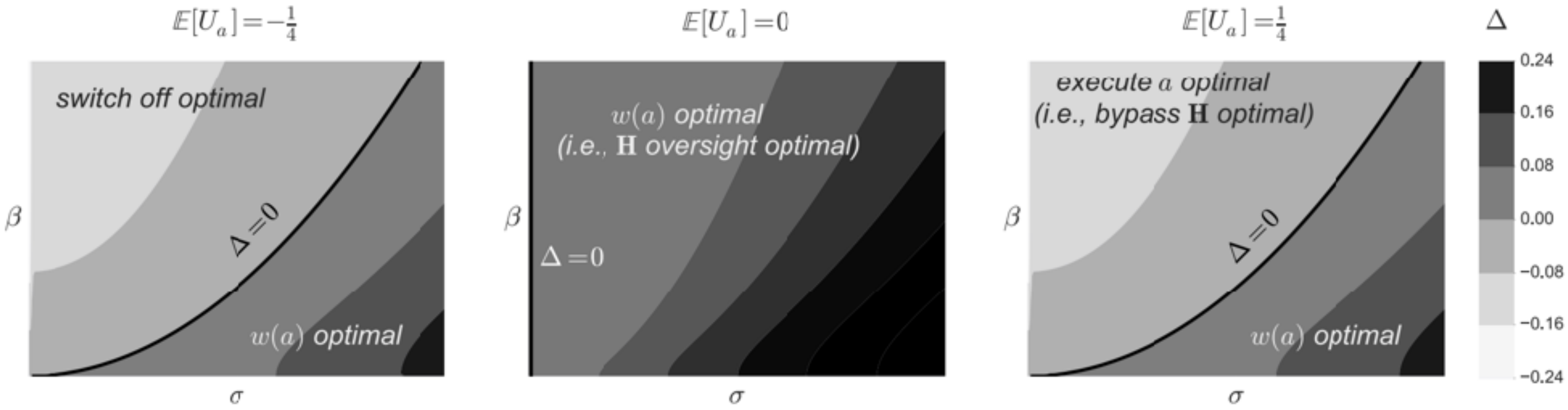
An autonomous car is driving an unaccompanied toddler to preschool. It would be irresponsible for the car to show the toddler a big red off switch.

Sub optimality (beta) here:

A noisily rational H models a human who occasionally makes the wrong decision in ‘unimportant situations.

Highlights the dependence of incentive on a trade off between R’s uncertainty and H’s sub optimality - What should R do ?

If H is an irrational/suboptimal actor

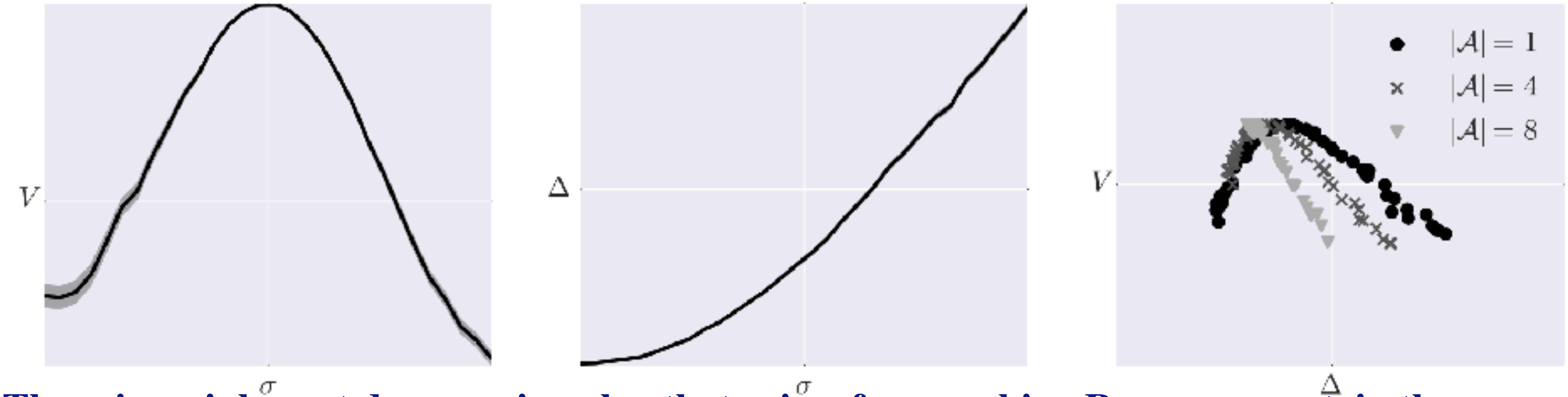


Analysis: R's belief over U_a, B^R, is a Gaussian distribution

- **Top-left:** Incentive is lower, where R is certain (Variance is low), H is very suboptimal (Beta is high)
- **Bottom-right:** Incentive is higher, where R is uncertain (Variance is high), H is near optimal (Beta is low)

If H is an irrational actor, then R may prefer switching itself off or executing a immediately rather than handing over the choice to H.

Loss in the utility of a system



There is an inherent decrease in value that arises from making R more uncertain than necessary. We measure this cost by considering the value in a modified off-switch game where R gets a noisy observation of H's preference.

- Left: The expected value V of the off-switch game as a function of the standard deviation in BR. V is maximized when is equal to the standard deviation that corresponds to the true Bayesian update.
- Middle: R's incentive to wait, as a function of . Together these show that, after a point, increasing , and hence increasing , leads to a decrease in V .
- Right: A scatter plot of V against . The different data series modify the number of potential actions R can choose among. If R has more choices, then obtaining a minimum value of will lead to a larger decrease in V .

Discussion notes