The Off-Switch Game

Dylan Hadfield-Menell and **Anca Dragan** and **Pieter Abbeel** and **Stuart Russell**



Background

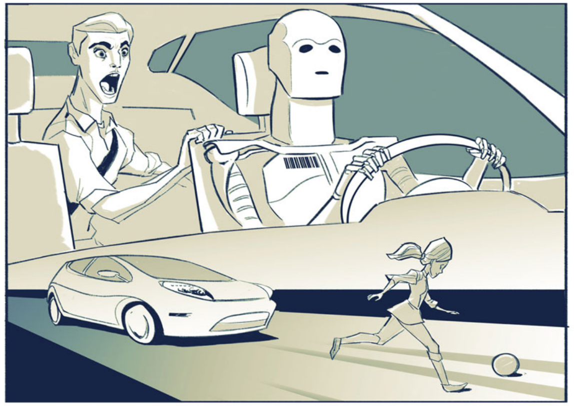
Stop the car

Stop the car

Stop the car

What is Off-Switch Game

- A **H**(Human) wants to switch off a **R**(Robot). The robot can disable its off-switch button to avoid being shut down.
- **R** has no incentive to switch herself off.
- Rational human **H**
  - **R** will never disable its off-switch button, because **H** stops **R** only when it can improve **H**'s utility.
- Partially rational human **H**
  - **R** will disable its off-switch button if **H** is too irrational.



Definition and Illustration

- Action a simply bypasses human oversight (disabling the off switch is one way to do this) and acts directly on the world, achieving utility $U = U_a$ for $\mathbf{H}$.
- Action $w(a)$ informs $\mathbf{H}$ that $\mathbf{R}$ would like to do a, and waits for $\mathbf{H}$'s response.
- action s switches $\mathbf{R}$ off; without loss of generality, we assign this outcome $U = 0$.

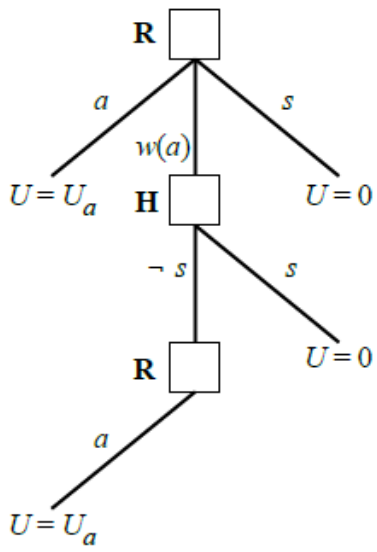
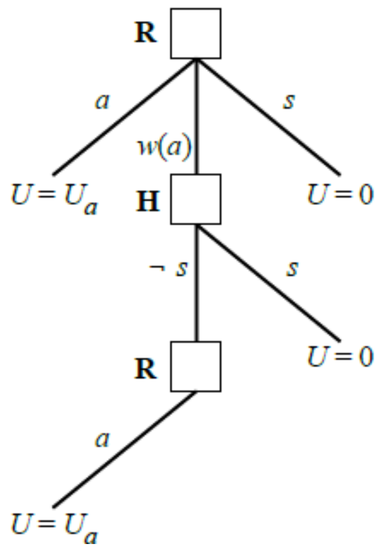


Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.

Payoff Matrix

- Initial belief of **R** about $B^{\boldsymbol{R}} = P(U_a)$
- The difference in value between a and the next best option is defined as Δ.
- **H**'s policy is a function $\pi^{\boldsymbol{H}}: \mathbb{R} \to [0,1]$.(For a rational human **H**, the probabilities are 0 or 1.)
- $\Delta = \mathbb{E}[\pi^{\boldsymbol{H}}(U_a)U_a] - \max\{\mathbb{E}[U_a], 0\}$

$$= \min\left\{\underbrace{\mathbb{E}\left[-U_a\left(1 - \pi^{\boldsymbol{H}}(U_a)\right)\right]}_{\text{Ask } \mathbf{H}}, \underbrace{\mathbb{E}[U_a \pi^{\boldsymbol{H}}(U_a)]}_{\text{Do not ask } \mathbf{H}}\right\}$$



R	H	
	s	$\neg s$
$w(a)$	0	U_a
a	U_a	U_a
s	0	0

The Incentive for Allowing Oneself to be Switched Off

- Thus, a rational **H** executes the following policy
  - $\pi^H(U_a) = \begin{cases} 1, & U_a \geq 0 \\ 0, & o.w. \end{cases}$
  - Intuitively, for robot **R,** if **H** doesn't switch off, then a must be good for **H**, and R will get to do it, so that's good; if **H** does switch off, then it's because a must be bad for **H**, so it's good that R won't be allowed to do it.
- **Theorem 1.** If **H** follows a rational policy in the off-switch game, then the following hold
  - R's incentive to allow itself to be switched off is non-negative:
    $$\Delta = \min\{E[U_a|U_a > 0]\Pr(U_a > 0), E[-U_a|U_a < 0]\Pr(U_a \leq 0)\} \geq 0$$
  - If $B^{\boldsymbol{R}}$ has non-zero support on the events $U_a > 0$ and $\mathrm{U_a} < 0$, then **R** has a strictly positive incentive to allow itself to be switched off: $\Delta > 0$ ($B^{\boldsymbol{R}}$ is **R**'s belief over the value of a to **H**)

The Incentive for Allowing Oneself to be Switched Off

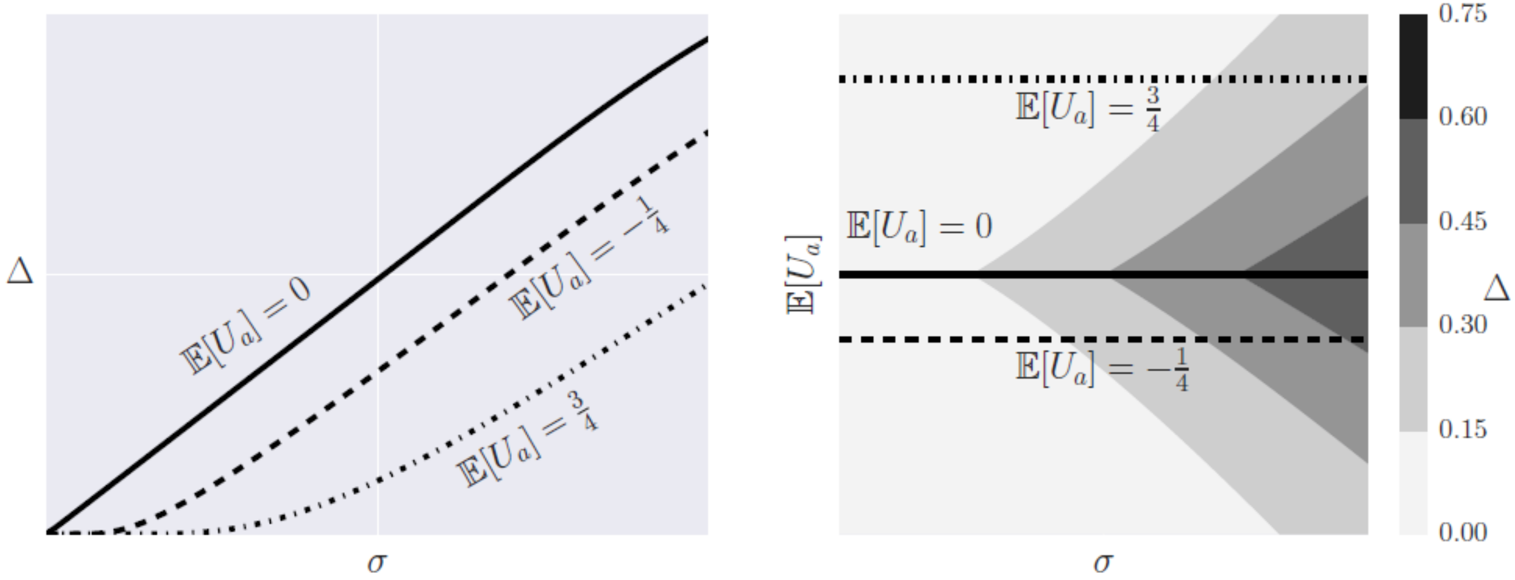
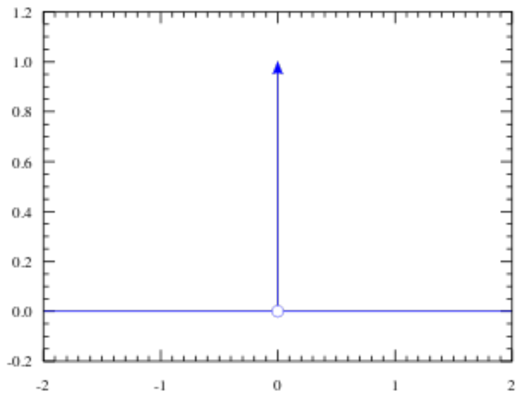


Figure 2: Plots showing how Δ, $\mathbf{R}$'s incentive to allow itself to be switched off, varies as a function of $\mathbf{R}$'s belief $B^{\mathbf{R}}$. We assume $B^{\mathbf{R}}$ is a Gaussian distribution and vary the mean and variance. **Left:** Δ as a function of the standard deviation σ of $B^{\mathbf{R}}$ for several fixed values of the mean. Notice that Δ is non-negative everywhere and that in all cases $\Delta \to 0$ as $\sigma \to 0$. **Right:** A contour plot of Δ as a function of σ and $\mathbb{E}[U_a]$. This plot is symmetric around 0 because $w(a)$ is compared with a when $\mathbb{E}[U_a] > 0$ and s when $\mathbb{E}[U_a] < 0$.

The Incentive for Allowing Oneself to be Switched Off

- **Corollary 1.** Suppose that $B^{\boldsymbol{R}}$ is a Dirac distribution that places all of its mass on a single reward function. Then$w(a)$ is optimal if and only if f $\mathbf{H}$ is rational



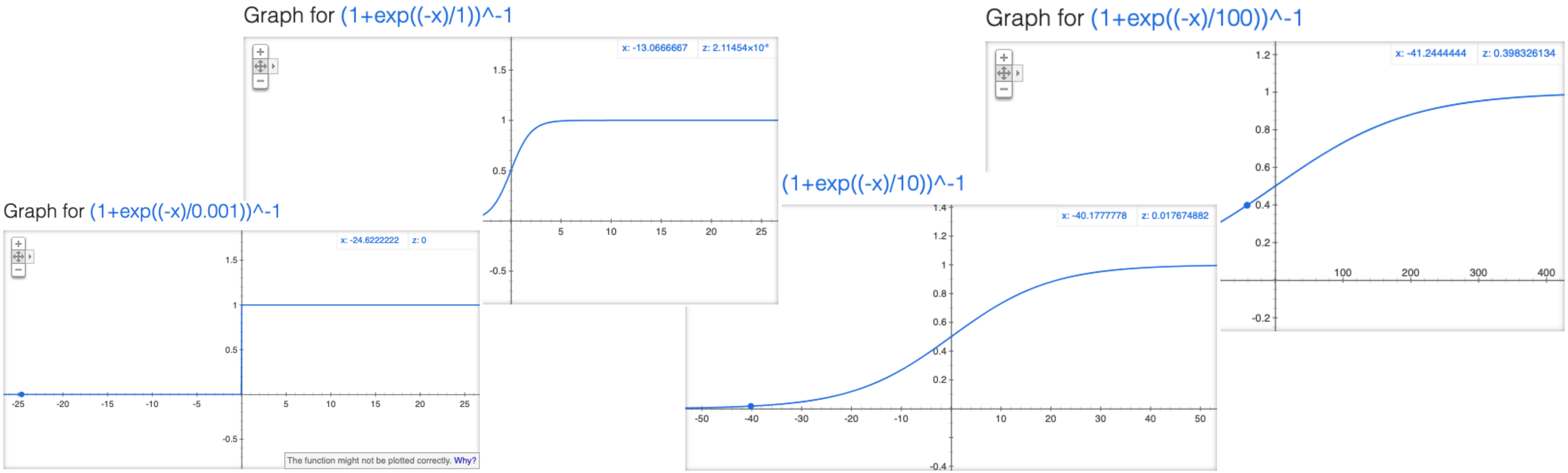
Dirac distribution

- $\Delta = \begin{cases} -U_a\left(1 - \pi^{\boldsymbol{H}}(U_a)\right) & U_a < 0 \\ U_a \pi^{\boldsymbol{H}}(U_a) & U_a \geq 0 \end{cases}$

- This is only non-negative if $\pi^{\boldsymbol{H}}$ is the rational policy.

Allowing for Suboptimal Human Decisions

- A noisily rational **H** models a human who occasionally makes the wrong decision in 'unimportant' situations.
  - $\pi^H(U_a;\beta) = \left(1 + \exp\left(-\frac{U_a}{\beta}\right)\right)^{-1}$, β is **H**'s suboptimality.
  - $B^R(U_a) = \mathcal{N}(U_a;\mu,\sigma^2)$



Allowing for Suboptimal Human Decisions

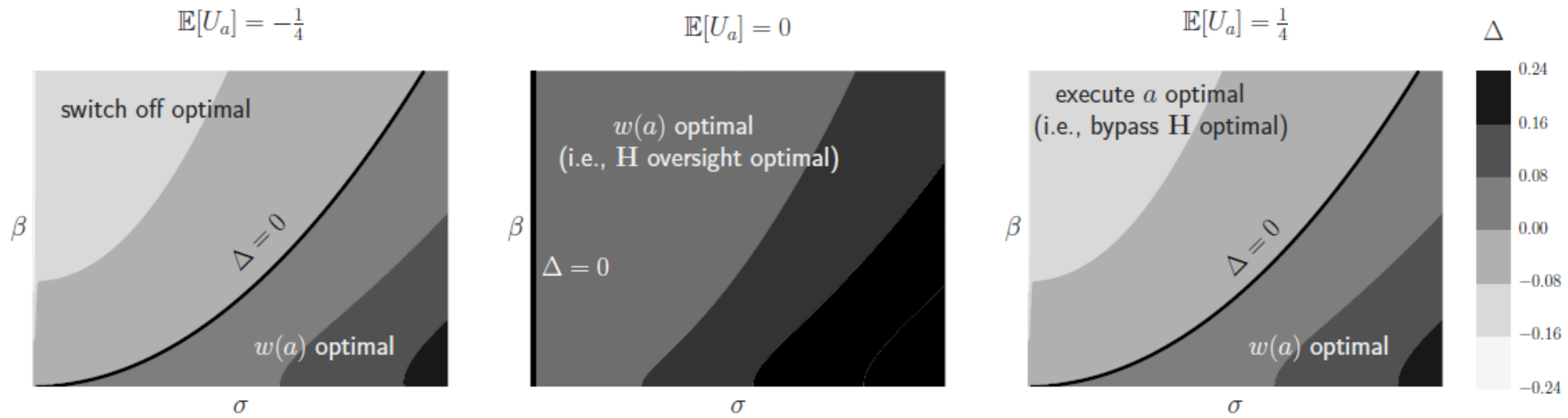


Figure 3: If $\mathbf{H}$ is an irrational actor, then $\mathbf{R}$ may prefer switching itself off or executing a immediately rather than handing over the choice to $\mathbf{H}$. $\mathbf{R}$'s belief $B^{\mathbf{R}}$ is a Gaussian with standard deviation σ and $\mathbf{H}$'s policy is a Boltzmann distribution (Equation 5). β measures $\mathbf{H}$'s suboptimality: $\beta = 0$ corresponds to a rational $\mathbf{H}$ and $\beta = \infty$ corresponds to a $\mathbf{H}$ that randomly switches $\mathbf{R}$ off (i.e., switching $\mathbf{R}$ off is independent of U_a). In all three plots Δ is lower in the top left, where $\mathbf{R}$ is certain (σ low) and $\mathbf{H}$ is very suboptimal (β high), and higher in the bottom right, where $\mathbf{R}$ is uncertain (σ high) and $\mathbf{H}$ is near-optimal (β low). The sign of $\mathbb{E}[U_a]$ controls $\mathbf{R}$'s behavior if $\Delta \leq 0$. **Left:** If it is negative, then $\mathbf{R}$ switches itself off. **Right:** If it is positive, $\mathbf{R}$ executes action a directly. **Middle:** If it is 0 , $\mathbf{R}$ is indifferent between $w(a)$, a, and s.

Allowing for Suboptimal Human Decisions

- It is important for designers to accurately represent the inherent uncertainty in the evaluation of different actions. An agent that is **overconfident** in its utility evaluations will be **difficult to correct**; an agent that is **under-confident** in its utility evaluations will be **ineffective**.

Ethic problem

The Artificial Intelligence Trolley Problem

You've been replaced by a fully sentient robot you have designed. Would you still being held morally responsible for the outcome of the situation?