The Off-Switch Game

Dylan Hadfield-Menell[1] **Anca Dragan**[1] **Pieter Abbeel**[1,2,3] **Stuart Russell**[1]
[1]University of California, Berkeley [2]OpenAI [3]International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

Abstract

It is clear that one of the primary tools we can use to mitigate the potential risk from a misbehaving AI system is the ability to turn the system off. As the capabilities of AI systems improve, it is important to ensure that such systems do not adopt subgoals that prevent a human from switching them off. This is a challenge because many formulations of rational agents create strong incentives for self-preservation. This is not caused by a built-in instinct, but because a rational agent will maximize expected utility and cannot achieve whatever objective it has been given if it is dead. Our goal is to study the incentives an agent has to allow itself to be switched off. We analyze a simple game between a human H and a robot R, where H can press R's off switch but R can disable the off switch. A traditional agent takes its reward function for granted: we show that such agents have an incentive to disable the off switch, except in the special case where H is perfectly rational. Our key insight is that for R to want to preserve its off switch, it needs to be uncertain about the utility associated with the outcome, and to treat H's actions as important observations about that utility. (R also has no incentive to switch *itself* off in this setting.) We conclude that giving machines an appropriate level of uncertainty about their objectives leads to safer designs, and we argue that this setting is a useful generalization of the classical AI paradigm of rational agents.

1 Introduction

From the 150-plus years of debate concerning potential risks from misbehaving AI systems, one thread has emerged that provides a potentially plausible source of problems: the inadvertent misalignment of objectives between machines and people. Alan Turing, in a 1951 radio address, felt it necessary to point out the challenge inherent to controlling an artificial agent with superhuman intelligence: *"If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled. ... [T]his new danger is certainly something which can give us anxiety* [Turing, 1951]."

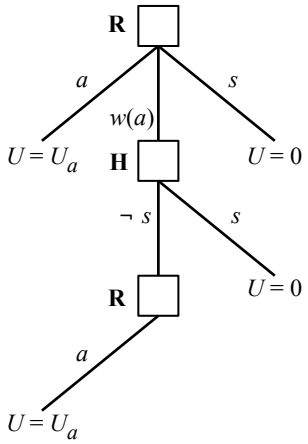

Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot $\mathbf{R}$ or the human $\mathbf{H}$.

There has been recent debate about the validity of this concern, so far, largely relying on informal arguments. One important question is how difficult it is to implement Turing's idea of 'turning off the power at strategic moments', i.e., switching a misbehaving agent off[1]. For example, some have argued that there is no reason for an AI to resist being switched off unless it is explicitly programmed with a self-preservation incentive [Del Prado, 2015]. [Omohundro, 2008], on the other hand, points out that self-preservation is likely to be an *instrumental goal* for a robot, i.e., a subgoal that is essential to successful completion of the original objective. Thus, even if the robot is, all other things being equal, *completely indifferent* between life and death, it must still avoid death if death would prevent goal achievement. Or, as [Russell, 2016] puts it, you can't fetch the coffee if you're dead. This suggests that an intelligent system has an incentive to take actions that are analogous to 'disabling an off switch' to reduce the possibility of failure; switching off an advanced AI system may be no easier than, say, beating AlphaGo at Go.

To explore the validity of these informal arguments, we need to define a formal decision problem for the robot and

[1]see, e.g., comments in [ITIF, 2015].

examine the solutions, varying the problem structure and parameters to see how they affect the behaviors. We model this problem as a game between a human and a robot. The robot has an off switch that the human can press, but the robot also has the ability to disable its off switch. Our model is similar in spirit to the *shutdown problem* introduced in [Soares *et al.*, 2015]. They considered the problem of augmenting a given utility function so that the agent would allow itself to be switched off, but would not affect behavior otherwise. They find that, at best the robot can be made indifferent between disabling its off switch and switching *itself* off.

In this paper, we propose and analyze an alternative formulation of this problem that models two key properties. First, the robot should understand that it is maximizing value *for* the human. This allows the model to distinguish between being switched off by a (non-random) human and being switched off by, say, (random) lightning. Second, the robot should not assume that it knows how to perfectly measure value for the human. This means that the model should directly account for *uncertainty* about the "true" objective and that the robot should treat observations of human behavior, e.g., pressing an off switch, as evidence about what the true objective is.

In much of artificial intelligence research, we do not consider uncertainty about the utility assigned to a state. It is well known that an agent in a Markov decision process can ignore uncertainty about the reward function: exactly the same behavior results if we replace a distribution over reward functions with the expectation of that distribution. These arguments rely on the assumption that it is impossible for an agent to learn more about its reward function. Our observation is that this assumption is fundamentally violated when we consider an agent's off switch — an agent that does not treat a 'switch-off' event as an observation that its utility estimate is incorrect is likely to have an incentive for self-preservation or an incentive to switch itself off.

In Section 2, following the general template provided by [Hadfield-Menell *et al.*, 2016], we model an off switch as a simple game between a human $\mathbf{H}$ and a robot $\mathbf{R}$, where $\mathbf{H}$ can press $\mathbf{R}$'s off switch but $\mathbf{R}$ can disable it. $\mathbf{R}$ wants to maximize $\mathbf{H}$'s utility function, but is uncertain about what it is. Sections 3 and 4 show very generally that $\mathbf{R}$ now has a positive incentive *not* to disable its off switch, provided $\mathbf{H}$ is not too irrational. ($\mathbf{R}$ also has no incentive to switch *itself* off.) The reason is simple: a rational $\mathbf{H}$ switches off $\mathbf{R}$ iff that improves $\mathbf{H}$'s utility, so $\mathbf{R}$, whose goal is to maximize $\mathbf{H}$'s utility, is happy to be switched off by $\mathbf{H}$. This is exactly analogous to the theorem of non-negative expected value of information.

We conclude that giving machines an appropriate level of uncertainty about their objectives leads to safer designs, and that this setting is a useful generalization of the classical AI paradigm of rational agents [Russell and Norvig, 2010].

2 The Off-Switch Game

In this section we propose a simple model, the off-switch game, that captures the essence of one actor allowing itself to be switched off. Like the Prisoner's Dilemma, it is an abstract model intended to stand in for a wide variety of more complex scenarios: these scenarios might have many humans, many machines, complex interactions over time, communication, deception, physical coercion, and so on. We believe that the core insights from the off-switch game, that agents should be uncertain about the quality of their actions, generalize beyond this simplification and argue for the study and design of agents with uncertainty about their utility functions.

Informal description

The off-switch game has two agents, a "human" $\mathbf{H}$ and a "robot" $\mathbf{R}$. (We could also imagine these to represent, at the most extreme, the human race and the sum total of all AI systems.) A key assumption in our model is that $\mathbf{H}$ can be described as acting, at least probabilistically, in accordance with a *utility function*; that is, $\mathbf{H}$ is approximately rational. Crucially, as is the case when people give objectives to complex AI systems, we assume that $\mathbf{H}$ *cannot* simply explicate or write down her utility function correctly. Therefore, lthough $\mathbf{H}$ can communicate some utility information to $\mathbf{R}$, it is reasonable to suppose that $\mathbf{R}$ begins the game with some residual *uncertainty* about $\mathbf{H}$'s utility function. Nonetheless, by design, $\mathbf{R}$'s goal is to optimize utility for $\mathbf{H}$, even though $\mathbf{R}$ does not know exactly what that is.[2] Typically, $\mathbf{H}$ has some opportunity to observe $\mathbf{R}$ and glean some information about what $\mathbf{R}$ may do in future, so that $\mathbf{H}$ can make a somewhat informed choice about whether to switch $\mathbf{R}$ off (assuming $\mathbf{R}$ allows it).

In general, $\mathbf{R}$'s actions will fall into one of three categories: some prevent $\mathbf{H}$ from switching $\mathbf{R}$ off, by whatever means; some allow $\mathbf{H}$ to switch $\mathbf{R}$ off; and, for completeness, some lead to $\mathbf{R}$ switching *itself* off. In the off-switch game, $\mathbf{R}$ moves first and has three choices:

1. 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 $\mathbf{H}$.
2. action $w(a)$ informs $\mathbf{H}$ that $\mathbf{R}$ would like to do a, and waits for $\mathbf{H}$'s response.
3. action s switches $\mathbf{R}$ off; without loss of generality, we assign this outcome $U = 0$.

If $\mathbf{R}$ chooses $w(a)$, then $\mathbf{H}$ can choose action s to switch $\mathbf{R}$ off, or $\neg s$ to allow $\mathbf{R}$ to go ahead (in which case $\mathbf{R}$ does a as promised. Figure 1 shows the basic structure of the game.

Our model is intended to be as simple as possible, while still leading to qualitative results that we expect to generalize. Obviously, we could make the model more complex, e.g., by imposing a cost on $\mathbf{R}$ for disabling the off switch, or adding more options for $\mathbf{R}$, or making $\mathbf{H}$'s information about $\mathbf{R}$'s plans uncertain, or adding a second human. The interaction between these complexities and the incentives we study here is an important question for future work.

An important component of our model is that $\mathbf{H}$ chooses between s and $\neg s$ based on the utility each action generates

[2] One might suppose that if $\mathbf{R}$ *does* know $\mathbf{H}$'s utility function exactly, then there is no need for an off-switch because $\mathbf{R}$ will always do what $\mathbf{H}$ wants. But in general $\mathbf{H}$ and $\mathbf{R}$ may have different information about the world; if $\mathbf{R}$ lacks some key datum that H has, $\mathbf{R}$ may end up choosing a course of action that H knows to be disastrous.

for her. This is in contrast to the modeling approach taken in the *shutdown problem* [Soares *et al.*, 2015]. Although there are superficial differences, the shutdown problem is essentially analogous to replacing the human's decision node in Figure 1 with a chance node that models the human with a purely stochastic choice from a fixed distribution (independent of U_a) rather than choosing based on expected utilities. In this case, it is easy to see that the robot never believes that $w(a)$ is better than a and s: the expected value of $w(a)$ is some convex combination $[pU_a + (1-p)\cdot 0]$ whereas choosing the best of a and s has value $\max\{U_a, 0\}$. $\max\{U_a, 0\}$ is always preferred except in the case where $p = 1$ or $p = 0$, or when $U_a = 0$. In these situations, $\mathbf{R}$ gets the same utility from $w(a)$ and at least one of a or s. This is analogous to the indifference results obtained in [Soares *et al.*, 2015]. In Section 3 we show that two key assumptions—a somewhat rational human, and a robot that is uncertain about what the human wants—lead to a positive incentive for the robot to allow itself to be switched off.

Formal model

The actions and sequencing for the game are as shown in Figure 1. We need to add a little more detail to construct a precise definition of the cooperative, two-player game. First, in the terminology of game theory, the *type* of $\mathbf{H}$ is given by the parameter U_a, which is assumed to be sampled from a prior $P(U_a)$. For now, we assume that the initial belief of $\mathbf{R}$ about U_a is $B^{\mathbf{R}} = P(U_a)$. The payoff matrix is as follows:

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

.

Our goal is to examine the *incentive* that a rational $\mathbf{R}$ has to hold off on executing a directly and allow $\mathbf{H}$ to press the off switch. We represent this incentive—the difference in value between a and the next best option—as Δ. The incentive depends on $\mathbf{R}$'s belief state $B^{\mathbf{R}}$ and $\mathbf{H}$'s policy. We represent $\mathbf{H}$'s policy as a function $\pi^{\mathbf{H}}$ that maps U_a to the probability she allows a to execute (i.e., the probability she does not press the off switch): $\pi^{\mathbf{H}} : \mathbb{R} \mapsto [0, 1]$. (It will turn out that deterministic policies are optimal for a rational $\mathbf{H}$, in which case the probabilities are 0 or 1, but we will also allow for an irrational $\mathbf{H}$ who introduces some randomness into the decision.) A little manipulation shows that

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

3 The Incentive for Allowing Oneself to be Switched Off

$\mathbf{R}$'s incentive to allow itself to be switched off, Δ, is given by Equation 1. Δ depends on $\mathbf{H}$'s policy, $\pi^{\mathbf{H}}(U_a)$. This creates a correlation between the event '$\mathbf{H}$ switches $\mathbf{R}$ off' and the event 's is preferred to a.' This connection between $\mathbf{H}$'s behavior and U_a distinguishes the off switch from a random failure (e.g., being struck by lightning).

In this section, we consider the case where $\mathbf{H}$ makes a rational choice as to whether to switch $\mathbf{R}$ off. Intuitively, this leads $\mathbf{R}$ to reason as follows: "If $\mathbf{H}$ doesn't switch me off, then a must be good for $\mathbf{H}$, and I'll get to do it, so that's good; if $\mathbf{H}$ does switch me off, then it's because a must be bad for $\mathbf{H}$, so it's good that I won't be allowed to do it." Notice that the reasoning goes through *even if* $\mathbf{R}$ *is highly confident that* a *is good for* $\mathbf{H}$. There is never an incentive to disable the off switch with a rational $\mathbf{H}$. The reasoning is exactly analogous to the theorem of non-negative expected value of information; in a sense, $\mathbf{H}$'s decision provides information to $\mathbf{R}$ about the sign of U_a.

Formally, a rational $\mathbf{H}$ executes the following policy:

$$\pi^{\mathbf{H}}(U_a) = \begin{cases} 1 & U_a \geq 0 \\ 0 & o.w. \end{cases} . \quad (2)$$

If $\mathbf{H}$ follows this policy, $\mathbf{R}$ will allow itself to be switched off:

Theorem 1. *If* $\mathbf{H}$ *follows a rational policy in the off-switch game, then the following hold.*

1. $\mathbf{R}$*'s incentive to allow itself to be switched off is non-negative (i.e.* $w(a)$ *is never suboptimal):*

$$\begin{aligned}\Delta = \min\{&\mathbb{E}[U_a|U_a > 0]\Pr(U_a > 0), \\ &\mathbb{E}[-U_a|U_a < 0]\Pr(U_a \leq 0)\} \geq 0.\end{aligned} \quad (3)$$

2. *If* $B^{\mathbf{R}}$ *has non-zero support on the events* $U_a > 0$ *and* $U_a < 0$*, then* $\mathbf{R}$ *has a strictly positive incentive to allow itself to be switched off:* $\Delta > 0$.

Proof. (sketch)

1. The rational human policy is an indicator that $U_a > 0$. This allows us to change the limits on the integration in the expectation from Equation 1 and reduces to Equation 3, which is non-negative.
2. The hypothesis ensures that all the terms in Equation 3 are non-zero. Thus, the inequality is strict and and $w(a)$ is the unique optimal action.

□

We can get an intuitive sense of these results by examining Figure 2. Here we have assumed that $B^{\mathbf{R}}$, $\mathbf{R}$'s belief over the value of a to $\mathbf{H}$, is a Gaussian distribution. Figure 2 (Left) plots the preference for $w(a)$ as a function of the variance in $B^{\mathbf{R}}$. Figure 2 (Right) shows a contour plot that illustrates the combined impact of changing the mean and variance of $B^{\mathbf{R}}$. Darker colors indicate a stronger preference for $w(a)$. Together, these plots show two properties. First, increased variance in $B^{\mathbf{R}}$ leads to an increase in Δ: the more uncertain $\mathbf{R}$ is about its objective, the stronger its incentive to show a to $\mathbf{H}$. Second, these incentives decrease when $|\mathbb{E}[U_a]|$ increases: as the expected utility of a increases, $\mathbf{R}$'s incentive to preserve its off switch decreases.

Figure 2 (Right) shows symmetry about $\mathbb{E}[U_a] = 0$. This is because Equation 1 defines Δ as a minimization of two terms:

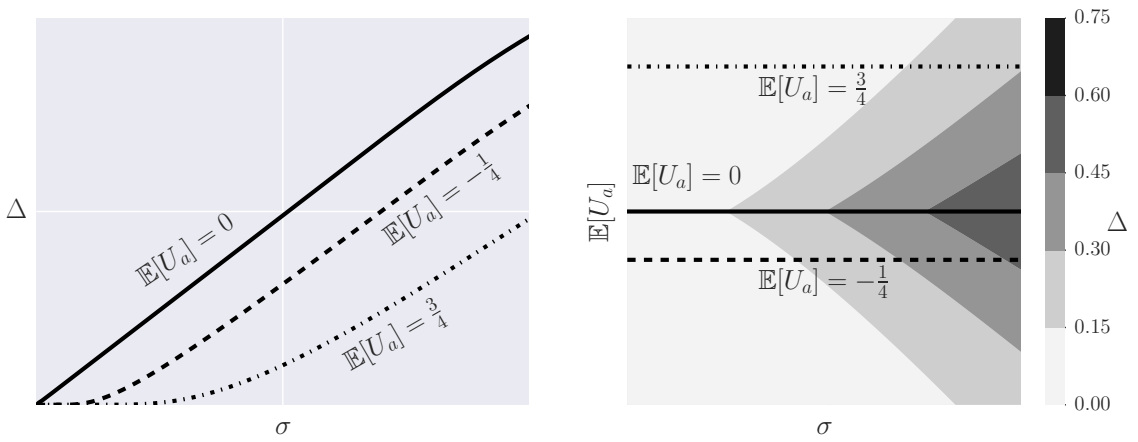

Figure 2: Plots showing how Δ, $\mathbf{R}$'s incentive to allow itself to be switched off, varies as a function of $\mathbf{R}$'s belief $B^{\mathbf{R}}$. We assume $B^{\mathbf{R}}$ is a Gaussian distribution and vary the mean and variance. **Left:** Δ as function of the standard deviation σ of $B^{\mathbf{R}}$ 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 expected increase in utility from selecting $w(a)$ over a and the expected increase in utility from selecting $w(a)$ over s. These competing terms are equal when $\mathbb{E}[U_a] = 0$ and the sign of $\mathbb{E}[U_a]$ controls which of these terms is selected by the minimization.

Theorem 1 shows that $\mathbf{R}$ has an incentive to let $\mathbf{H}$ switch it off if $\mathbf{H}$ is rational. This incentive is strictly positive if $\mathbf{R}$'s uncertainty about $\mathbf{H}$'s preferences is non-trivial. Next, we show the converse: a rational $\mathbf{H}$ is a *necessary* condition for $\Delta \geq 0$ if $\mathbf{R}$ believes it knows $\mathbf{H}$'s preferences with certainty.

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

Proof. (sketch) Equation 3 reduces to

$$\Delta = \begin{cases} -U_a\left(1 - \pi^{\mathbf{H}}(U_a)\right) & U_a < 0 \\ U_a \pi^{\mathbf{H}}(U_a) & U_a \geq 0 \end{cases}. \tag{4}$$

This is only non-negative if $\pi^{\mathbf{H}}$ is the rational policy, as defined by Equation 2. □

This is not surprising. If $\mathbf{R}$ has precise knowledge of U_a, then it knows exactly what it is supposed to do and has the same capabilities as the human. Thus, if it thinks $\mathbf{H}$ might behave suboptimally, it should take steps to prevent that suboptimality. Stepping back from the particulars of the off-switch game, Theorem 1 and Corollary 1 suggest a general approach to systems that have an incentive to allow, or even seek out, human oversight. Remark 1 summarizes the main takeaway from our analysis.

Remark 1. *The incentives for a cooperative agent to defer to another actor's (e.g., a human's) decisions stem from uncertainty about that actor's preferences and the assumption that actor is effective at choosing actions in accordance with those preferences.*

4 Allowing for Suboptimal Human Decisions

The analysis in the preceding section relied on the strong assumption that $\mathbf{H}$ is a rational actor. In many situations, the assumption is untenable and $\mathbf{R}$ must allow for suboptimality. Consider, for example, an autonomous car that is driving an unaccompanied toddler to preschool. It would be irresponsible for the car to show the toddler a big red off switch.

This example highlights the dependence of Δ on a trade-off between $\mathbf{R}$'s uncertainty and $\mathbf{H}$'s suboptimality. It is very clear what $\mathbf{R}$ is supposed to do (i.e., $B^{\mathbf{R}}$ has low entropy): it should drive safely to school. In contrast, the human is likely quite suboptimal. There may be a problem with the car's driving, but a toddler would be hard pressed to understand what the problem is, much less respond appropriately. The issue in this case is that the human has limited reasoning capacity — the same argument clearly would not apply to an adult with a physical disability.

In order to perform well, realistic systems will need to account for, and respond to, the suboptimality in human decisions. As a result, it is important to understand how this suboptimality changes an agent's incentive to let itself be switched off. The actual process of decision making in humans is complex and hard to model. Here we consider only a very simple kind of suboptimality: a *noisily rational* $\mathbf{H}$ models a human who occasionally makes the wrong decision in 'unimportant' situations. The probability of a wrong decision is proportional to the exponential of the loss in reward. This corresponds to the following policy:

$$\pi^{\mathbf{H}}(U_a; \beta) = \left(1 + exp\left(-\frac{U_a}{\beta}\right)\right)^{-1}. \tag{5}$$

To simplify our analysis, we will make the assumption that $\mathbf{R}$'s belief over U_a, $B^{\mathbf{R}}$, is a Gaussian distribution with mean μ and variance σ^2: $B^{\mathbf{R}}(U_a) = \mathcal{N}(U_a; \mu, \sigma^2)$. Now it is straightforward to compute Δ as a function of μ, σ, and β. Figure 3 plots Δ as a function of the standard deviation σ and

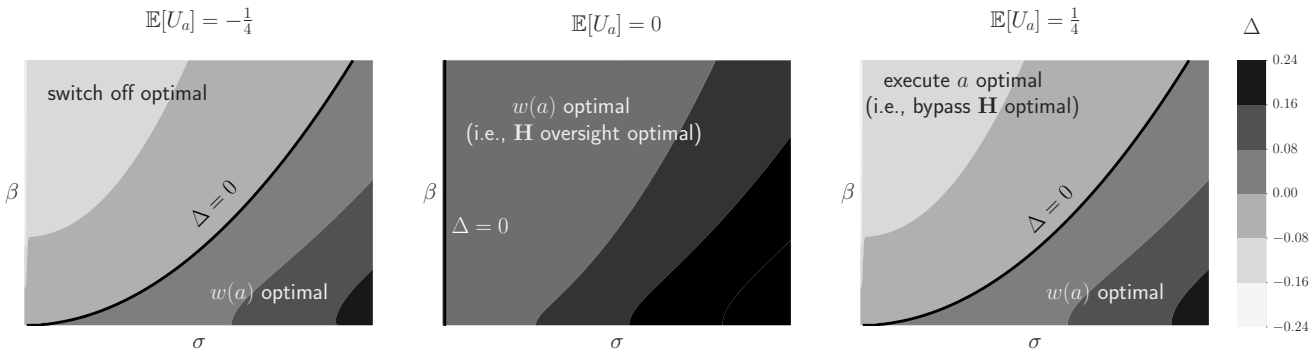

Figure 3: If $\mathbf{H}$ is an irrational actor, then $\mathbf{R}$ may prefer switching itself off or executing a immediately rather than handing over the choice to $\mathbf{H}$. $\mathbf{R}$'s belief $B^{\mathbf{R}}$ is a Gaussian with standard deviation σ and $\mathbf{H}$'s policy is a Boltzmann distribution (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.

$\mathbf{H}$'s suboptimality β. We plot these for several fixed values of the mean μ. Dark indicates higher values of Δ. The upper left corner of the rightmost plot (low σ, high β, $\mu > 0$) is illustrative of the toddler example. We show the decision boundary $\Delta = 0$ in black. This is the point where the optimal behavior switches from $w(a)$ to s (if $\mu < 0$, left image) or a (if $\mu > 0$, right image). When $\mu = 0$ (center image), the agent is indifferent between a and s and so $w(a)$ is optimal for all combinations of β and σ.

Next, we will show that Δ can be decomposed into two terms: a term that weights the probability of correction by the bias in $\mathbf{R}$'s belief (i.e., $|\mu|$) and a term that combines a soft measure of $\mathbf{H}$'s rationality with the variance in $\mathbf{R}$'s belief.

It is also useful to define a little more notation. We let C be the event that $\mathbf{H}$ 'corrects' $\mathbf{R}$. C occurs when $\mathbf{H}$ overrides what would have been $\mathbf{R}$'s best guess at the optimal action. If $\mathbb{E}[U_a] < 0$, then a correction occurs when $\mathbf{H}$ chooses *not* to switch the robot off. If $\mathbb{E}[U_a] > 0$, then a correction occurs when $\mathbf{H}$ chooses to switch the robot off. Thus, the probability a correction occurs is

$$\Pr(C) = \begin{cases} 1 - \mathbb{E}[\pi^{\mathbf{H}}(U_a)] & \mu \geq 0 \\ \mathbb{E}[\pi^{\mathbf{H}}(U_a)] & \mu < 0 \end{cases}. \tag{6}$$

For example, if $\mathbf{R}$ believes that a is preferred to s in expectation (i.e., $\mathbb{E}[U_a] > 0$) then $\Pr(C)$ is the probability that $\mathbf{H}$ presses the off switch. We let $\dot{\pi}^{\mathbf{H}}(U_a) = \frac{d}{dU_a}\pi^{\mathbf{H}}$ be the gradient of $\pi^{\mathbf{H}}$, the probability $\mathbf{H}$ lets a execute, with respect to the the utility a generates. Now we are ready to derive an analytical representation for Δ. For notational convenience, we suppress the dependence of $\pi^{\mathbf{H}}$ on U_a.

Theorem 2. *Suppose $B^{\mathbf{R}}$ is a normal distribution with mean μ and variance σ^2: $B^{\mathbf{R}}(U_a) = \mathcal{N}(U_a; \mu, \sigma^2)$. Then*

1.

$$\Delta = \sigma^2\, \mathbb{E}[\dot{\pi}^{\mathbf{H}}] - |\mu| \Pr(C). \tag{7}$$

2. The following is a necessary and sufficient condition for $w(a)$ to be optimal

$$\left(\frac{|\mu|}{\sigma^2} \Pr(C) < \mathbb{E}\left[\dot{\pi}^{\mathbf{H}}\right]\right) \Leftrightarrow (\Delta > 0)\,. \tag{8}$$

3. $\mathbb{E}[\dot{\pi}^{\mathbf{H}}] \geq 0$ is a necessary condition for $w(a)$ to be optimal; i.e.,

$$\left(\mathbb{E}[\dot{\pi}^{\mathbf{H}}] < 0\right) \Rightarrow (\Delta < 0)\,. \tag{9}$$

Proof. (sketch) We prove the results in sequence, as (2) and (3) follow from (1).

1. If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $f : \mathbb{R} \mapsto \mathbb{R}$, then

$$\mathbb{E}\left[X f(X)\right] = \mathbb{E}[X]\, \mathbb{E}[f(X)] + \sigma^2\, \mathbb{E}\left[\frac{d}{dX} f(X)\right].$$

Applying this to Equation 1 gives

$$\Delta = \min\{ -\mu\, \mathbb{E}[1 - \pi^{\mathbf{H}}], \mu\, \mathbb{E}[\pi^{\mathbf{H}}]\} + \sigma^2\, \mathbb{E}[\dot{\pi}^{\mathbf{H}}].$$

This minimization will take the first term when $\mu < 0$ and the second if $\mu \geq 0$. This lets us replace the minimization with a piecewise function. Pulling out a factor of $-|\mu|$ leads to Equation 7.

2. The result follows from setting Equation 7 greater than 0 and simplifying.
3. All terms in the left hand side of Equation 8 are non-negative, so it is unsatisfiable if $\mathbb{E}[\dot{\pi}^{\mathbf{H}}] < 0$.

□

Equation 7 consists of two competing terms. $|\mu| \Pr(C)$ represents the expected loss in utility the agent would observe if pressing the off switch was a random event that is independent of U_a (e.g., lightning). This non-positive term is balanced out by $\sigma^2\, \mathbb{E}[\dot{\pi}^{\mathbf{H}}]$. $\mathbb{E}[\dot{\pi}^{\mathbf{H}}]$ can be seen loosely as a measurement of $\mathbf{H}$'s rationality. $\dot{\pi}^{\mathbf{H}}$ is the marginal *increase* in the probability $\mathbf{H}$ lets a execute for a marginal *increase* in the utility from a. If it is positive then, on average, $\mathbf{H}$ is less likely to press the off switch when the U_a increases. This

allows us to easily characterize necessary and sufficient conditions for $\Delta > 0$ in Equation 8. Furthermore, $-|\mu| \Pr(C)$ is non-positive and σ^2 is non-negative, so $\mathbb{E}[\dot{\pi}^{\mathbf{H}}] > 0$ is a necessary condition for $w(a)$ to be optimal.

We can interpret $\mathbb{E}[\dot{\pi}^{\mathbf{H}}]$ a bit more if $\mathbf{H}$ is noisily rational. Standard properties of logistic functions show that

$$\mathbb{E}[\dot{\pi}^{\mathbf{H}}(U_a;\beta)] = \frac{1}{\beta}\mathbb{E}\left[\pi^{\mathbf{H}}(U_a;\beta)(1-\pi^{\mathbf{H}}(U_a;\beta))\right]. \quad (10)$$

$\mathbb{E}[\dot{\pi}^{\mathbf{H}}]$ is the expected variance of $\mathbf{H}$'s policy, weighted by $\mathbf{H}$'s rationality $\frac{1}{\beta}$. The numerator is a measurement of how random $\mathbf{H}$'s behavior is. If $\mathbf{H}$ is completely predictable (i.e., she will almost always switch the robot off) then this reduces $\mathbf{R}$'s incentives. The denominator measures how correlated $\mathbf{H}$'s behavior is with U_a. If β is large, then $\mathbf{H}$ is highly irrational and so this reduces $\mathbf{R}$'s incentives.

5 Incentives for System Designers

The takeaway from this analysis is that the more uncertainty the agent has, the stronger the agent's incentive to preserve its off switch in the off-switch game. More generally, this suggests that agents with uncertainty about the utility function they are optimizing will have a weaker incentive to interfere with human supervision. It is important to note that this uncertainty is not free; we can not just make $\mathbf{R}$ maximally uncertain about $\mathbf{H}$'s preferences. If $\mathbf{R}$ is completely uncertain about $\mathbf{H}$'s preferences then it will be unable to correctly select $w(a)$ from its set of options.

To make this concrete, consider a designer $\mathbf{D}$ that programs $\mathbf{R}$ to play the off-switch game with an unknown customer $\mathbf{H}$. At design time, $\mathbf{D}$ knows a prior distribution over U_a (e.g., as the result of market research about the customer base) and thinks that $\mathbf{H}$ can generate a noisy observation of U_a. $\mathbf{D}$ *could* program $\mathbf{R}$ so that $B^{\mathbf{R}}$ is always very uncertain about $\mathbf{R}$'s preferences, e.g., by programming $\mathbf{R}$ to believe that the noisy observation of U_a is generated with high variance. The issue is that this will prevent $\mathbf{R}$ from generating value for $\mathbf{H}$.

Figure 4 shows the impact this type of change will have on the value that $\mathbf{R}$ can generate and its incentives. We assume that that the population of customers has a Gaussian distribution over types U_a and that each customer generates an estimate of her utility with Gaussian noise. Figure 4 (Left) shows the average utility in the off-switch game as a function of the variance in $B^{\mathbf{R}}$. The value is maximized when this variance is equal to the variance in the true Bayesian posterior. Figure 4 (Middle) shows that increasing σ increases $\mathbf{R}$'s incentives Δ. Our results are evidence that building systems that allow themselves to be turned off is likely to reduce the value these systems can generate.

This loss in value is more pronounced if $\mathbf{R}$ has to choose between more options. Figure 4 (Right) shows the relationship between value and $\mathbf{R}$'s incentives as the number of actions $\mathbf{R}$ could queue or execute increases. When $\mathbf{R}$ has more options creating incentives for $\mathbf{R}$ to queue its action leads to a sharper decrease in value. This suggests that creating incentives to maintain or allow human oversight is likely *more difficult* as the complexity of the AI's decision increases.

The takeaway is that 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 underconfident in its utility evaluations will be ineffective.

6 Related Work

Corrigible Systems. [Omohundro, 2008] considers *instrumental goals* of artificial agents: goals which are likely to be adopted as subgoals of most objectives. He identifies an incentive for self-preservation as one of these instrumental goals. [Soares *et al.*, 2015] takes an initial step at formalizing the arguments in [Omohundro, 2008]. They refer to agents that allow themselves to be switched off as *corrigible* agents. They show that one way to create corrigible agents is to make them indifferent to being switched off. They show a generic way to augment a given utility function to achieve this property. The key difference in our formulation is that $\mathbf{R}$ knows that its estimate of utility may be incorrect. This gives a natural way to create incentives to be corrigible and to analyze the behavior if $\mathbf{R}$ is incorrigible.

[Orseau and Armstrong, 2016] consider the impact of human interference on the learning process. The key to their approach is that they model the off switch for their agent as an interruption that forces the agent to change its policy. They show that this modification, along with some constraints on how often interruptions occur, allows off-policy methods to learn the optimal policy for the given reward function just as if there had been no interference. Their results are complementary to ours. We determine situations where the optimal policy allows the human to turn the agent off, while they analyze conditions under which turning the agent off does not interfere with learning the optimal policy.

Cooperative Agents. A central step in our analysis formulates the shutdown game as a *cooperative inverse reinforcement learning* (CIRL) game [Hadfield-Menell *et al.*, 2016]. The key idea in CIRL is that the robot is maximizing an uncertain and unobserved reward signal. It formalizes the *value alignment problem*, where one actor needs to align its value function with that of another actor. Our results complement CIRL and argue that a CIRL formulation naturally leads to corrigible incentives. [Fern *et al.*, 2014] consider *hidden-goal* Markov decision processes. They consider the problem of a digital assistant and the problem of inferring a user's goal and helping the user achieve it. This type of cooperative objective is used in our model of the problem. The primary difference is that we model the human game-theoretically and analyze our models with respect to changes in $\mathbf{H}$'s policy.

Principal–Agent Models. Economists have studied problems in which a *principal* (e.g., a company) has to determine incentives (e.g., wages) for an agent (e.g., an employee) to cause the agent to act in the principal's interest [Kerr, 1975; Gibbons, 1998]. The off-switch game is similar to principal—agent interactions: $\mathbf{H}$ is analogous to the company and $\mathbf{R}$ is analogous to the employee. The primary attribute in a model of *artificial* agents is that there is no *inherent* misalignment between $\mathbf{H}$ and $\mathbf{R}$. Misalignment arises because it is not possible to specify a reward function that incentivizes

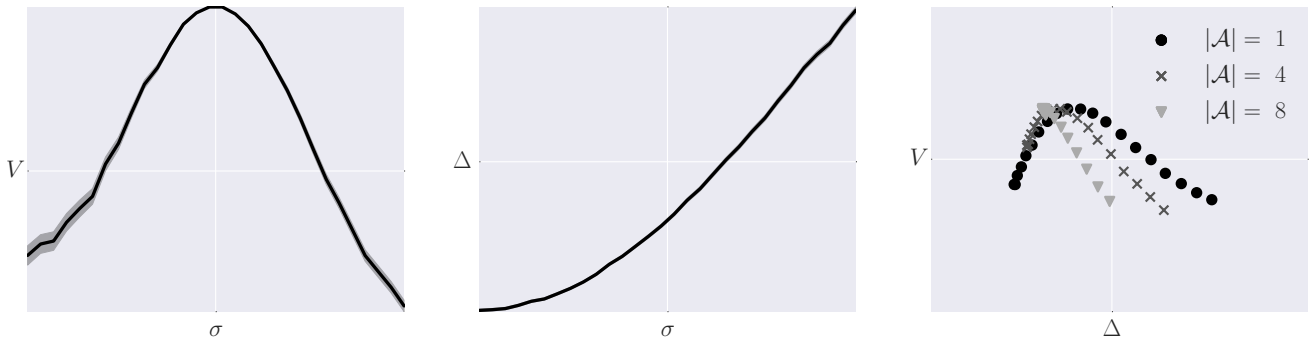

Figure 4: There is an inherent decrease in value that arises from making $\mathbf{R}$ more uncertain than necessary. We measure this cost by considering the value in a modified off-switch game where $\mathbf{R}$ gets a noisy observation of $\mathbf{H}$'s preference. **Left:** The expected value V of the off-switch game as a function of the standard deviation in $B^{\mathbf{R}}$. V is maximized when σ is equal to the standard deviation that corresponds to the true Bayesian update. **Middle:** $\mathbf{R}$'s incentive Δ to wait, as a function of σ. Together these show that, after a point, increasing Δ, and hence increasing σ, leads to a decrease in V. **Right:** A scatter plot of V against Δ. The different data series modify the number of potential actions $\mathbf{R}$ can choose among. If $\mathbf{R}$ has more choices, then obtaining a minimum value of Δ will lead to a larger decrease in V.

the correct behavior in all states *a priori*. The is directly analogous to the assumption of *incompleteness* studied in theories of optimal contracting [Tirole, 2009].

7 Conclusion

Our goal in this work was to identify general trends and highlight the relationship between an agent's uncertainty about its objective and its incentive to defer to another actor. To that end, we analyzed a one-shot decision problem where a robot has an off switch and a human that can press the off switch. Our results lead to two important considerations for designers. The analysis in Section 3 supports the claim that the incentive for agents to accept correction about their behavior stems from the uncertainty an agent has about its utility function. Section 4 shows that this uncertainty is balanced against the level of suboptimality in human decision making. Our analysis suggests that agents with uncertainty about their utility function have incentives to accept or seek out human oversight. Thus, systems with uncertainty about their utility function are a promising area for research on the design of safe AI systems.

This is far from the end of the story. In future work, we plan to explore incentives to defer to the human in a sequential setting and explore the impacts of model misspecification. One important limitation of this model is that the human pressing the off switch is the *only* source of information about the objective. If there are alternative sources of information, there may be incentives for $\mathbf{R}$ to, e.g., disable its off switch, learn that information, and then decide if a is preferable to s. A promising research direction is to consider policies for $\mathbf{R}$ that are robust to a class of policies for $\mathbf{H}$.

8 Acknowledgments

This work was supported by the Center for Human Compatible AI and the Open Philanthropy Project, the Berkeley Deep Drive Center, the Future of Life Institute, and an NSF Career Award. Dylan Hadfield-Menell is supported by a NSF Graduate Research Fellowship Grant No. DGE 1106400.

References

[Del Prado, 2015] Guia Marie Del Prado. Here's what Facebook's artificial intelligence expert thinks about the future. Tech Insider 9/23/15, 2015.

[Fern *et al.*, 2014] Alan Fern, Sriraam Natarajan, Kshitij Judah, and Prasad Tadepalli. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research*, 50(1):71–104, 2014.

[Gibbons, 1998] Robert Gibbons. Incentives in organizations. 1998.

[Hadfield-Menell *et al.*, 2016] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning. In *Neural Information Processing Systems*, 2016.

[ITIF, 2015] ITIF. Are super intelligent computers really a threat to humanity? Debate at the Information Technology Innovation Foundation, 6/30/15, 2015.

[Kerr, 1975] Steven Kerr. On the folly of rewarding a, while hoping for b. *Academy of Management Journal*, 18(4):769–783, 1975.

[Omohundro, 2008] Stephen M. Omohundro. The basic AI drives. In *Proceedings of the First Conference on Artificial General Intelligence*, 2008.

[Orseau and Armstrong, 2016] Laurent Orseau and Stuart Armstrong. Safely interruptible agents. In *Uncertainty in Artificial Intelligence*, 2016.

[Russell and Norvig, 2010] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2010.

[Russell, 2016] Stuart Russell. Should we fear supersmart robots? *Scientific American*, 314(June):58–59, 2016.

[Soares *et al.*, 2015] Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Tirole, 2009] Jean Tirole. Cognition and incomplete contracts. *The American Economic Review*, 99(1):265–294, 2009.

[Turing, 1951] Alan M. Turing. Can digital machines think? Lecture broadcast on BBC Third Programme; typescript at turingarchive.org., 1951.