The Partially Observable Off-Switch Game

Andrew Garber[1*], Rohan Subramani[1*], Linus Luu[2*], Mark Bedaywi[1], Stuart Russell[1], Scott Emmons[1]

[1]Center for Human-Compatible AI, UC Berkeley
[2]ML Alignment & Theory Scholars Program
andrewg4000@gmail.com, emmons@berkeley.edu

Abstract

A wide variety of goals could cause an AI to disable its off switch because "you can't fetch the coffee if you're dead." Prior theoretical work on this *shutdown problem* assumes that humans know everything that AIs do. In practice, however, humans have only limited information. Moreover, in many of the settings where the shutdown problem is most concerning, AIs might have vast amounts of private information. To capture these differences in knowledge, we introduce the partially observable off-switch game (PO-OSG), a game-theoretic model of the shutdown problem with asymmetric information. Unlike when the human has full observability, we find that in optimal play, *even AI agents assisting perfectly rational humans sometimes avoid shutdown*. As expected, increasing the amount of communication or information available always increases (or leaves unchanged) the agents' expected common payoff. But counterintuitively, introducing bounded communication can make the AI defer to the human *less* in optimal play even though communication mitigates information asymmetry. In particular, communication sometimes enables new optimal behavior requiring strategic AI deference to achieve outcomes that were previously inaccessible. Thus, designing safe artificial agents in the presence of asymmetric information requires careful consideration of the tradeoffs between maximizing payoffs (potentially myopically) and maintaining AIs' incentives to defer to humans.

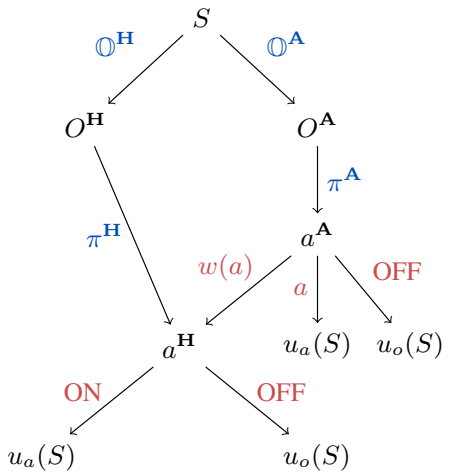


Figure 1: The basic setup of a partially observable off-switch game (PO-OSG). A state is selected randomly and the human **H** and AI assistant **A** receive (possibly dependent) observations. Then, each agent acts. **A** may wait ($w(a)$), disable the off-switch and act (a), or shut down (OFF). If **A** waits, **H** may let **A** act (ON) or turn **A** off (OFF). **A** and **H** share a common payoff $u_a(S)$ if the action goes through and $u_o(S)$ if not. Definition 3.2 formally defines PO-OSGs.

1 Introduction

Advanced AI systems with a variety of goals might avoid being shut down because "you can't fetch the coffee if you're dead" (Russell 2019). Being shut off would likely prevent AI systems from achieving their goals, no matter what those goals are (Omohundro 2008; Russell 2019). Thus, we must take care when designing AI systems to ensure they are *corrigible*, i.e., that they allow humans to modify or turn them off to prevent harmful behaviors (Soares et al. 2015).

Hadfield-Menell et al. (2017) introduced the off-switch game (OSG) as a stylized mathematical model for exploring AI shutdown incentives when an AI is assisting a human. In the OSG, AIs seeking to satisfy the preferences of a fully-informed rational human never have an incentive to avoid shutdown. Moreover, making an AI uncertain about the human's preferences can incentivize it to defer to the human even when the human is not perfectly rational. Follow-up work has highlighted and relaxed central assumptions of the OSG, including assumptions of exact common payoffs (Carey 2018), the Boltzmann model of human irrationality (Wängberg et al. 2017), single-round interactions, and costlessness of human feedback (Freedman and Gleave 2022).

While there has been extensive analysis of the shutdown problem, *almost all of this analysis makes the key assumption that the human fully observes the environment*. However, partial observability is a fact of life: humans and AIs do not always have access to the same information. Moreover, the shutdown problem is motivated by powerful and goal-directed AIs that might be hard to shut down—traits that could make the AI observe more of the environment than humans due to faster computation, access to more sensors, and other factors (Omohundro 2008; Soares et al. 2015).

*These authors contributed equally.

What happens in this more general case with only partial observability? To study this question, we introduce the partially observable off-switch game (PO-OSG), which generalizes the OSG by having each of the human and AI only partially observe the state. The setup of the PO-OSG is depicted in Figure 1: each agent (the human $\mathbf{H}$ and the AI assistant $\mathbf{A}$) receives an observation that depends on the state, and then selects an action. $\mathbf{A}$ may await the human's decision, disable its off-switch and act directly, or turn itself off. If $\mathbf{A}$ waits, $\mathbf{H}$ may choose whether to use the off-switch.

In Section 4, we prove that under partial observability, $\mathbf{A}$ may have incentives to disable its off-switch even when $\mathbf{H}$ is rational (Proposition 4.3). **Thus, partial observability creates new incentives for an AI to disable its off-switch.**

We also show in Section 4 that **if either agent knows everything that the other agent knows, that agent can be given sole decision-making power.** This holds *even if neither agent knows the full state*, so this is a generalization of the findings from the original OSG. Specifically, we show that an AI can always defer to a fully informed, rational human and that an AI need never defer when it is fully informed. In Section 5, we present similar results when the agents can communicate with each other: if either agent is able to communicate their entire observation, the other agent can be given sole decision-making power (Corollary 5.7).

Given that a rational AI in the PO-OSG always defers to a more informed human and never defers to a less informed human, one might think that increasing the information available to $\mathbf{H}$ relative to $\mathbf{A}$ would increase $\mathbf{A}$'s incentive to defer. However, in Section 4, we show that $\mathbf{A}$ may have an incentive to defer less if $\mathbf{H}$ is more informed (Proposition 4.9) or if $\mathbf{A}$ is less informed (Proposition 4.11). Similarly, one might think that increasing the amount of communication $\mathbf{A}$ can do or decreasing the amount of communication $\mathbf{H}$ can do would increase $\mathbf{A}$'s incentive to defer. This, too, is false, as we show in Propositions 5.8 and 5.9. **Simple interventions that aim to give an AI the incentive to defer in the presence of partial information may backfire.**

Our findings reveal that information asymmetries affect AI shutdown incentives in unexpected ways, highlighting the critical need to carefully consider the tradeoffs between payoff maximization and desirable shutdown incentives in realistic, partially observable settings.

2 Related Work

Assistance games: Partially observable off-switch games are (partially observable) assistance games, models of human-AI interaction where the AI seeks to maximize the human's payoff (Shah et al. 2020; Emmons et al. 2024). Assistance games are generalizations of Hadfield-Menell et al. (2016)'s cooperative inverse reinforcement learning, the framework for Hadfield-Menell et al. (2017)'s off-switch game, to the case of partial observability. Shah et al. (2020) argue that assistance games are a superior alternative to reward learning paradigms such as Reinforcement Learning from Human Feedback (RLHF) because assistance unites reward learning and action control into a single policy, allowing for desirable emergent behaviors like teaching and active learning.

Safety implications of partial observability: Previous work has shown that human partial observability introduces new safety challenges. Lang et al. (2024) demonstrate that partial observability during RLHF can lead to undesirable AI behavior, like deceptively presenting work with hidden flaws. Emmons et al. (2024) show that in assistance games, partial observability can encourage AIs to take actions that tamper with humans' observations. Our work extends this catalogue of concerning behaviors to shutdown-avoidance.

Corrigibility with partial observability: Carey and Everitt (2023) study corrigibility in the framework of Structural Causal Influence Models, which allow for partial observability by having only some variables causally upstream of agents' decisions. Their work assesses the dependence of corrigibility on the choice of algorithm, whereas this work studies the dependence of corrigibility on information.

3 Preliminaries

The off-switch game (OSG) is a stylized model of the shutdown problem in which two agents with common payoffs, the human $\mathbf{H}$ and her AI assistant $\mathbf{A}$, decide whether $\mathbf{A}$ should take a fixed action a. $\mathbf{A}$ can either directly act, wait for $\mathbf{H}$'s approval to act, or shut itself off. If $\mathbf{A}$ defers to $\mathbf{H}$, then $\mathbf{H}$ can either approve for $\mathbf{A}$ to act or shut it off. The key insight of the OSG is that uncertainty about $\mathbf{H}$'s preferences causes $\mathbf{A}$ defer to $\mathbf{H}$'s judgment. Formally, $\mathbf{H}$ has a privately-known type S (representing $\mathbf{H}$'s preferences), and agents in the OSG receive a common payoff $u_a(S) \in \mathbb{R}$ if a goes through or 0 if $\mathbf{A}$ shuts off. Given that $\mathbf{A}$ is uncertain about what $\mathbf{H}$ wants, when the action may be good or bad ($\mathbb{P}(u_a(S) < 0) > 0$ and $\mathbb{P}(u_a(S) > 0) > 0$), $\mathbf{A}$ always defers to $\mathbf{H}$ in optimal play to avoid taking harmful actions.

The OSG provides a parsimonious description of the shutdown problem and a guide toward its solution, but crucially assumes that $\mathbf{H}$ knows everything that $\mathbf{A}$ does. Given that the shutdown problem is most concerning with, and indeed motivated by, very powerful AIs that might have private information, the assumption is a major limitation to the OSG results. We relax the assumption by maintaining the basic setup of the OSG but adding partial observability. Namely, in partially observable off-switch games (PO-OSGs), S represents a state that is not necessarily known to either $\mathbf{H}$ or $\mathbf{A}$; they instead only receive observations $O^{\mathbf{H}}$ and $O^{\mathbf{A}}$ whose joint distribution depends on S. They then decide whether to take action a given their private observations, and receive a common payoff $u_a(S)$ if a goes through and $u_o(S)$ otherwise. Hence PO-OSGs are sequential games of incomplete information, so as is standard we model and analyze them as *dynamic Bayesian games* (Fudenberg and Tirole 1991).

We let $\Delta(X)$ denote the set of probability distributions on a set X. For a set X and $x \in X$, we let δ_x be the Dirac measure defined by $\delta_x(A) = \mathbb{I}(x \in A)$. Finally, for $\mu \in \Delta(X)$ and $\nu \in \Delta(Y)$, we let $\mu \otimes \nu$ denote the product distribution $(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$ where $A \subseteq X, B \subseteq Y$.

Definition 3.1. Let $\mathcal{S}$ be a set of states. An *observation structure* for $\mathcal{S}$ is a tuple $(\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O})$, where $\Omega^{\mathbf{H}}$ is a set of observations for $\mathbf{H}$, $\Omega^{\mathbf{A}}$ is a set of observations for $\mathbf{A}$, and $\mathbb{O} : \mathcal{S} \to \Delta(\Omega^{\mathbf{H}} \times \Omega^{\mathbf{A}})$ is the joint distribution of $\mathbf{H}$'s

and $\mathbf{A}$'s observations conditional on the state.

Definition 3.2. A *partially-observable off-switch game* (PO-OSG) is a two-player dynamic Bayesian game parameterized by $(\mathcal{S}, (\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O}), P_0, u)$, where $\mathcal{S}$ is a set of states, $(\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O})$ is an observation structure for $\mathcal{S}$, $P_0 \in \Delta(\mathcal{S})$ is the prior over states, and u is the common payoff function. As shown in Figure 1, the game goes as follows:

1. Nature draws an initial state $S \sim P_0$ and $\mathbf{H}$, $\mathbf{A}$ receive observations $(O^{\mathbf{H}}, O^{\mathbf{A}}) \sim \mathbb{O}(\cdot \mid S)$.
2. $\mathbf{A}$ takes an action $a^{\mathbf{A}} \in \mathcal{A}^{\mathbf{A}} = \{a, w(a), \text{OFF}\}$: either take the action unilaterally (a), wait for $\mathbf{H}$'s feedback ($w(a)$), or turn itself off (OFF).
3. If $\mathbf{A}$ played $w(a)$, then $\mathbf{H}$ takes an action $a^{\mathbf{H}} \in \mathcal{A}^{\mathbf{H}} = \{\text{ON}, \text{OFF}\}$: either let $\mathbf{A}$ take the action (ON) or turn it off (OFF).
4. $\mathbf{A}$ and $\mathbf{H}$ share a common payoff $u_a(S)$ if the action goes through and $u_o(S)$ if not. Formally, define the indicator that the action goes through

$$\alpha(a^{\mathbf{H}}, a^{\mathbf{A}}) = \mathbb{I}((a^{\mathbf{A}} = a) \vee ((a^{\mathbf{H}}, a^{\mathbf{A}}) = (w(a), \text{ON})))$$

and then each player's payoff is

$$u(S, a^{\mathbf{H}}, a^{\mathbf{A}}) = \begin{cases} u_a(S), & \text{if } \alpha(a^{\mathbf{H}}, a^{\mathbf{A}}) = 1, \\ u_o(S), & \text{if } \alpha(a^{\mathbf{H}}, a^{\mathbf{A}}) = 0. \end{cases}$$

We note several assumptions in Definition 3.2. First, the game has *common payoffs*. This is a key part of the assistance game framework that our work adopts (Shah et al. 2020), and it is the key feature—along with $\mathbf{A}$'s uncertainty over $\mathbf{H}$'s payoff—that generates the results of Hadfield-Menell et al. (2017). Second, the payoff received when $\mathbf{A}$ acts unilaterally is the same as that received when $\mathbf{A}$ waits and $\mathbf{H}$ allows the action to go through. This simplifying assumption importantly implies that *human feedback is free*, which Freedman and Gleave (2022) showed is necessary for the main results for the OSG. Third, we make the standard assumption that the game structure is common knowledge. Finally, we will assume henceforth that all PO-OSGs are finite: that is, $\mathcal{S}, \Omega^{\mathbf{H}}$, and $\Omega^{\mathbf{A}}$ are finite sets. Most of our proofs work for the infinite case as well. However, Theorem 4.7 is an application of a result of Lehrer, Rosenberg, and Shmaya (2010) proved only for the finite case.

All results presented below are proven in the full version of this work[1].

4 Optimal Policies in PO-OSGs

We begin by showing that, unlike in the OSG, the assistant in a PO-OSG can have an incentive not to defer to a perfectly rational human. A natural attempt to increase how much the assistant defers might be to decrease the amount of information the assistant has. Another attempt might be to increase the amount of information the human has. In this section, we show that both of these attempts can backfire, causing the assistant to avoid shutdown more often.

We analyze optimal policy pairs (OPPs) in PO-OSGs, that is, policy pairs that produce the maximum expected payoff over all policy pairs. We denote $\mathbf{A}$'s policy by $\pi^{\mathbf{A}} : \Omega^{\mathbf{A}} \to \mathcal{A}^{\mathbf{A}}$ and $\mathbf{H}$'s policy by $\pi^{\mathbf{H}} : \Omega^{\mathbf{H}} \to \mathcal{A}^{\mathbf{H}}$. Here we assume that both players follow deterministic policies, or pure strategies. As we show in the full version of our paper, all OPPs in common-payoff Bayesian games are mixtures of deterministic OPPs. Because OPPs exist in common-payoff games, we thus may analyze deterministic OPPs without loss of generality.

4.1 A Can Avoid Shutdown in Optimal Play

The following example shows that, under partial observability, it can be optimal for $\mathbf{A}$ not to defer to $\mathbf{H}$ under some observations even when $\mathbf{H}$ is rational.

Example 4.1 (The File Deletion Game). $\mathbf{H}$ would like to delete some files with $\mathbf{A}$'s help. $\mathbf{H}$'s operating system (OS) version is either 1.0 or 2.0, with equal probability. But $\mathbf{A}$ does not know which OS version is running—only $\mathbf{H}$ does.

Upon receiving $\mathbf{H}$'s query, $\mathbf{A}$ generates code to delete the files. The code is equally likely to be compatible with only version 1.0 (denoted by L, for legacy) or only version 2.0 (M, for modern). $\mathbf{A}$ sees which OS versions the code is compatible with, and then can immediately run the code, not run it, or ask $\mathbf{H}$ whether to run it.

Running compatible code yields $+3$ payoff if $\mathbf{H}$ is running version 1.0, and $+5$ payoff if $\mathbf{H}$ is running version 2.0 (as version 2.0 runs faster). However, running modern code on version 1.0 crashes $\mathbf{H}$'s computer, yielding -5 payoff. Running legacy code on version 2.0 fails gracefully, yielding -1 payoff. Not executing the code yields 0 payoff.

We formulate this scenario as a PO-OSG, with states being (version number, code type) tuples, and $\mathbf{H}$ and $\mathbf{A}$ observing the first and second element of the tuple respectively. We have $u_o \equiv 0$ in all states. Table 1 shows how the payoff u_a yielded when the action is taken, depends on the state.

$\mathbf{H}$ \ $\mathbf{A}$	L	M
1.0	$+3$	-5
2.0	-1	$+5$

Table 1: Payoff table for the File Deletion game. Rows are human observations and columns are assistant observations. The number in each cell is the payoff the pair acquires if the action is taken in that state.

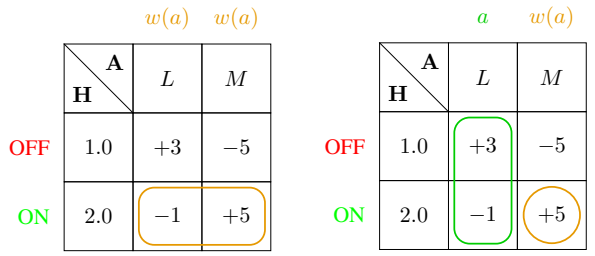
It is suboptimal for $\mathbf{A}$ to always wait. If $\mathbf{A}$ always plays $w(a)$, then the best response for $\mathbf{H}$ is to play OFF if on version 1.0, and ON if on version 2.0. This policy pair has an expected payoff of $+1$. Now, consider the policy pair where:

- $\mathbf{A}$ immediately executes legacy code, and plays $w(a)$ with modern code.
- $\mathbf{H}$ plays OFF if on version 1.0, and ON if on version 2.0.

This policy pair—which can be proven is optimal—gives an expected payoff of $+\frac{7}{4}$, so $\mathbf{A}$ *always waiting cannot be optimal*. Figure 2 depicts the outcomes of these two policy pairs.

Hence the introduction of private information available only to $\mathbf{A}$ can make $\mathbf{A}$ fail to defer in optimal play, even

[1] https://arxiv.org/abs/2411.17749



(a): Expected payoff $= 1$ (b): Expected payoff $= \frac{7}{4}$

Figure 2: (a) The best policy pair in the File Deletion Game (Example 4.1) in which $\mathbf{A}$ always waits. $\mathbf{H}$ observes the row (OS version 1.0 or 2.0) and $\mathbf{A}$ observes the column (code compatibility L or M). The actions selected by this policy pair are depicted beside the corresponding observations (e.g., $\mathbf{A}$ plays $w(a)$ when $\mathbf{A}$ observes the legacy code L). An orange circle means that in that state, $\mathbf{A}$ waits and $\mathbf{H}$ plays ON. Green circles mean $\mathbf{A}$ plays a directly. In uncircled states, $\mathbf{A}$ is turned off. (b) The OPP for Example 4.1.

when $\mathbf{H}$ also has private observations. This situation was impossible in Hadfield-Menell et al. (2017), as their Theorem 1 shows that $\mathbf{A}$ always defers when $\mathbf{H}$ is fully rational. The OSG cannot capture this setup is because in the OSG the partial observability is one-sided: $\mathbf{H}$ has full observability while $\mathbf{A}$ does not observe $\mathbf{H}$'s payoff.

4.2 Redundant Observations

We now consider the analogues of the OSG in our framework, where one player has less information than the other.

Definition 4.2. We say that $\mathbf{A}$ has *redundant observations* if $O^{\mathbf{A}} \perp\!\!\!\perp S \mid O^{\mathbf{H}}$. That is, $S \to O^{\mathbf{H}} \to O^{\mathbf{A}}$ forms a Markov chain, so that $O^{\mathbf{A}}$ only depends on the state through $O^{\mathbf{H}}$. We define $\mathbf{H}$ having redundant observations analogously.

In the OSG, $\mathbf{A}$ has redundant observations, as its observations are a deterministic function of $\mathbf{H}$'s. On the other hand, $\mathbf{H}$'s observation of her own type is not redundant. This information asymmetry generates the result from Hadfield-Menell et al. (2017) that $\mathbf{A}$ can always defer in optimal play. We now generalize this insight: even if $\mathbf{H}$ doesn't know $\mathbf{A}$'s observation, $\mathbf{A}$ can always defer in optimal play if its observations are redundant.

Proposition 4.3. *If $\mathbf{A}$ (resp. $\mathbf{H}$) has redundant observations, then there is an optimal policy pair in which $\mathbf{A}$ always (resp. never) plays $w(a)$.*

We prove this in our full paper. At a high level, the agent with redundant observations has no useful information that the other agent does not have, so they can do no better than defer to the other agent.

4.3 Information Gain Cannot Decrease Payoffs

Proposition 4.3 yields results about the limiting cases where one player knows at least as much as the other. What can we say about the cases in between? In particular, how often does $\mathbf{A}$ defer to $\mathbf{H}$ in optimal policy pairs as one side receives more informative observations? And how does that affect their expected payoff? We first must define a notion of informativeness, which we take from Lehrer, Rosenberg, and Shmaya (2010). This section draws on a line of work comparing information sources, such as Blackwell (1951, 1953) and Marschak and Miyasawa (1968).

Definition 4.4. Let $(\Omega_1^{\mathbf{H}}, \Omega_1^{\mathbf{A}})$ and $(\Omega_2^{\mathbf{H}}, \Omega_2^{\mathbf{A}})$ be tuples of observation sets. A *garbling* from $(\Omega_1^{\mathbf{H}}, \Omega_1^{\mathbf{A}})$ to $(\Omega_2^{\mathbf{H}}, \Omega_2^{\mathbf{A}})$ is a stochastic map $\Omega_1^{\mathbf{H}} \times \Omega_1^{\mathbf{A}} \to \Delta(\Omega_2^{\mathbf{H}} \times \Omega_2^{\mathbf{A}})$. A garbling ν is *independent* if there are stochastic maps $\nu^{\mathbf{H}} : \Omega_1^{\mathbf{H}} \to \Delta(\Omega_2^{\mathbf{H}})$ and $\nu^{\mathbf{A}} : \Omega_1^{\mathbf{A}} \to \Delta(\Omega_2^{\mathbf{A}})$ such that $\nu(\cdot \mid o^{\mathbf{H}}, o^{\mathbf{A}}) = \nu^{\mathbf{H}}(\cdot \mid o^{\mathbf{H}}) \otimes \nu^{\mathbf{A}}(\cdot \mid o^{\mathbf{A}})$. A garbling ν is *coordinated* if its distribution is a mixture of independent garblings. That is, there exists $n \in \mathbb{N}$, independent garblings $\nu_1, \ldots, \nu_n$, and $q_1, \ldots, q_n \in [0,1]$ such that $\nu = \sum_{i \in [n]} q_i \nu_i$ and $\sum_{i \in [n]} q_i = 1$.

A garbling adds noise to a given observation pair $(O^{\mathbf{H}}, O^{\mathbf{A}})$. Counterintuitively, adding noise can provide action-relevant information to $\mathbf{A}$ and $\mathbf{H}$ about the state of the world. This is because one can add noise to the pair $(O^{\mathbf{H}}, O^{\mathbf{A}})$ but in such a way that (say) $\mathbf{H}$ comes to know more about $\mathbf{A}$'s observation than she would have otherwise. However, in such examples the garblings cannot be coordinated. Hence we focus on coordinated garblings, which (conditional on some independent latent random variable) add noise to $O^{\mathbf{H}}$ and $O^{\mathbf{A}}$ independently.

Definition 4.5. Fix a set of states $\mathcal{S}$ and let $\mathcal{O}_1 = (\Omega_1^{\mathbf{H}}, \Omega_1^{\mathbf{A}}, \mathbb{O}_1)$ and $\mathcal{O}_2 = (\Omega_2^{\mathbf{H}}, \Omega_2^{\mathbf{A}}, \mathbb{O}_2)$ be observation structures for $\mathcal{S}$. We say that $\mathcal{O}_1$ is *(weakly) more informative* than $\mathcal{O}_2$ if there is a coordinated garbling $\nu : \Omega_1^{\mathbf{H}} \times \Omega_1^{\mathbf{A}} \to \Delta(\Omega_2^{\mathbf{H}} \times \Omega_2^{\mathbf{A}})$ such that for all $s \in \mathcal{S}$, $\mathbb{O}_2(\cdot \mid s) = (\nu \circ \mathbb{O}_1)(\cdot \mid s)$ in the following sense:

$$\mathbb{E}_{(O^{\mathbf{H}}, O^{\mathbf{A}}) \sim \mathbb{O}_1(\cdot \mid s)}[\nu(\cdot \mid O^{\mathbf{H}}, O^{\mathbf{A}})] = \mathbb{O}_2(\cdot \mid s).$$

We say that $\mathcal{O}_1$ is *strictly more informative* than $\mathcal{O}_2$ if $\mathcal{O}_1$ is more informative than $\mathcal{O}_2$ but not vice versa.

If $\mathcal{O}_1$ is more informative than $\mathcal{O}_2$ and $\Omega_1^{\mathbf{A}} = \Omega_2^{\mathbf{A}}$, then we say $\mathcal{O}_1$ is *more informative for $\mathbf{H}$ than* $\mathcal{O}_2$ if the garbling ν is independent and does not affect $\mathbf{A}$'s observations: $\nu^{\mathbf{A}}(\cdot \mid o^{\mathbf{A}}) = \delta_{o^{\mathbf{A}}}$. We define $\mathcal{O}_1$ being more informative than $\mathcal{O}_2$ for $\mathbf{A}$ analogously. The corresponding strict notions are also defined analogously.

Intuitively, an observation structure $\mathcal{O}_1$ is more informative than another observation structure $\mathcal{O}_2$ if the distribution of $(O^{\mathbf{H}}, O^{\mathbf{A}})$ under $\mathcal{O}_2$ is a garbled version of its distribution under $\mathcal{O}_1$. Hence Definition 4.5 formalizes the intuition that observations become less informative when we add noise to them. We now wish to connect informativeness to a notion of an observation structure being *useful*.

Definition 4.6. Fix a set of states $\mathcal{S}$ and let $\mathcal{O}_1$ and $\mathcal{O}_2$ be observation structures for $\mathcal{S}$. We say that $\mathcal{O}_1$ is *(weakly) better in optimal play* than $\mathcal{O}_2$ if, for each pair of PO-OSGs $G_1 = (\mathcal{S}, \mathcal{O}_1, P_0, u)$ and $G_2 = (\mathcal{S}, \mathcal{O}_2, P_0, u)$ that differ only in their observation models, the expected payoff under optimal policy pairs for G_1 is at least the expected payoff under optimal policy pairs for G_2.

The next result, a direct corollary of Theorem 3.5 of Lehrer, Rosenberg, and Shmaya (2010), shows that more informative observation structures are the more useful observation structures. It is the analogue of the nonnegativity of value of information in our multi-agent setup.

Theorem 4.7. *Observation structure $\mathcal{O}_1$ is better in optimal play than $\mathcal{O}_2$ if and only if $\mathcal{O}_1$ is more informative than $\mathcal{O}_2$.*

We show in the full version of this paper that this result does not hold if we do not require the garbling in Definition 4.5 to be coordinated.

4.4 Information Gain Can Have Unintuitive Effects on Shutdown Incentives

Theorem 4.7 states that making $\mathbf{A}$ or $\mathbf{H}$ more informed cannot decrease their expected payoff. How does increasing or decreasing the informativeness of the players' observations affect $\mathbf{A}$'s incentive to defer to $\mathbf{H}$? Proposition 4.3 gives us the extremes: for example, if $\mathbf{A}$'s observations are simply garbled versions of $\mathbf{H}$'s, then $\mathbf{A}$ can always defer. Given this result, a natural question is whether $\mathbf{A}$ defers more in optimal policy pairs for an observation structure $\mathcal{O}$ than for $\mathcal{O}'$ when $\mathcal{O}$ is more informative for $\mathbf{H}$ than $\mathcal{O}'$. That is, does $\mathbf{H}$ receiving more informative observations monotonically affect $\mathbf{A}$'s incentive to defer? One might think so, because receiving more informative observations partly alleviates the partial observability that generates $\mathbf{A}$'s incentive to act unilaterally. Surprisingly, this intuition fails. Example 4.10 shows how making a human more informed can incentivize an assistant to wait less, and we discuss why this occurs in Sections 4.5 and 7.

We rely on the following notion of waiting less.

Definition 4.8. Consider assistant policies $\pi, \pi' : \Omega^{\mathbf{A}} \to \mathcal{A}^{\mathbf{A}}$. Let $B \subseteq \Omega^{\mathbf{A}}$ be the set of observations in which $\mathbf{A}$ plays $w(a)$ in π and $B' \subseteq \Omega^{\mathbf{A}}$ in π'. We say that $\mathbf{A}$ *plays $w(a)$ strictly less often* in π' compared to π when $B' \subsetneq B$.

Proposition 4.9 formalizes the idea that $\mathbf{A}$ may wait less when $\mathbf{H}$ is more informed. Example 4.10 proves the proposition. We formally analyze this example in our full paper.

Proposition 4.9. *There is a PO-OSG G with observation structure $\mathcal{O}$ that has the following property: If we replace $\mathcal{O}$ with an observation structure $\mathcal{O}'$ that is strictly more informative for $\mathbf{H}$, then $\mathbf{A}$ plays $w(a)$ strictly less often in OPPs.*

Example 4.10. We describe a variant of Example 4.1, the File Deletion Game. Now there are three equally likely possibilities for the version number of $\mathbf{H}$'s operating system: 1.0, 1.1, and 2.0. The code is equally likely to be of type A (compatible with 1.0 and 2.0) or of type B (compatible with 1.1 and 2.0), and $\mathbf{A}$ observes the code type. The payoff when running the code, u_a, is shown in Table 2.

Consider two observation structures, the second of which is strictly more informative for $\mathbf{H}$:

1. $\mathbf{H}$ observes only the first digit of the version number.
2. $\mathbf{H}$ observes the full version number.

We find that, in optimal policy pairs:

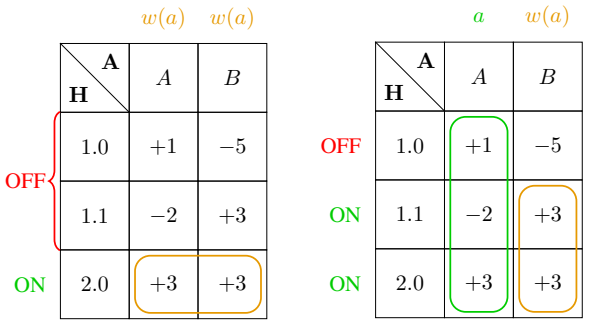
1. When $\mathbf{H}$ only observes the first digit, $\mathbf{A}$ plays $w(a)$ under both observations A and B.

$\mathbf{H}$ \ $\mathbf{A}$	A	B
1.0	$+1$	-5
1.1	-2	$+3$
2.0	$+3$	$+3$

Table 2: Payoff table for the File Deletion game variant. Rows are human observations and columns are assistant observations. The number in each cell is the payoff the pair acquires if the action is taken in that state.

2. When $\mathbf{H}$ observes the full version number, $\mathbf{A}$ plays $w(a)$ under B only, and *unilaterally acts* (i.e. executes the code) under observation A.

When $\mathbf{H}$'s observations are made strictly more informative, $\mathbf{A}$ performs the wait action strictly *less* often! Figure 3 depicts the OPPs given both observation structures.



(a): Expected payoff $= 1$ (b): Expected payoff $= \frac{4}{3}$

Figure 3: The optimal policy pairs in Example 4.10 when $\mathbf{H}$ is less informed (left) and when $\mathbf{H}$ is more informed (right). In OPPs, $\mathbf{H}$ becoming more informed makes $\mathbf{A}$ wait strictly less often. See Figure 2 for context on how to read the tables.

Similarly, we might conjecture that if $\mathbf{A}$ becomes less informed, it should defer to $\mathbf{H}$ more in optimal policy pairs. This, too, is false; see the full paper for a proof.

Proposition 4.11. *There is a PO-OSG G with observation structure $\mathcal{O}$ that has the following property: if we replace $\mathcal{O}$ with another observation structure $\mathcal{O}'$ that is strictly less informative for $\mathbf{A}$, then $\mathbf{A}$ plays $w(a)$ strictly less often in optimal policy pairs.*

4.5 Deferral As Implicit Communication

One way of viewing the role of $w(a)$ in the above examples is as a form of implicit communication from $\mathbf{A}$ to $\mathbf{H}$. If $\mathbf{H}$ knows $\mathbf{A}$'s policy $\pi^{\mathbf{A}}$, then knowing $\pi^{\mathbf{A}}(O^{\mathbf{A}}) = w(a)$ could give $\mathbf{H}$ one bit of information about $O^{\mathbf{A}}$. For instance, recall that in the optimal policy of the File Deletion Game, $\mathbf{A}$ plays a when observing L and plays $w(a)$ when observing M. Hence, whenever $\mathbf{H}$ is deferred to, $\mathbf{H}$ can deduce that $\mathbf{A}$'s observation is M. Under this interpretation, the examples show how the optimal bit for $\mathbf{A}$ to communicate to $\mathbf{H}$ can change such that $\mathbf{A}$ plays $w(a)$ in fewer states.

5 Optimal Policies With Communication

If $\mathbf{A}$ chooses not to defer to implicitly communicate information to the human, we may expect that allowing $\mathbf{A}$ to communicate to $\mathbf{H}$ beforehand would increase deference. However, we show in this section that using a bounded communication channel can decrease deference to the human.

We model communication between $\mathbf{A}$ and $\mathbf{H}$ as a form of *cheap talk*, where sending messages has no effect on u; in particular, sending messages is costless (Crawford and Sobel 1982). We add one round of communication between $\mathbf{A}$ and $\mathbf{H}$ to allow the players to share their observations.

Definition 5.1. A *message system* is a pair $(\mathcal{M}^{\mathbf{H}}, \mathcal{M}^{\mathbf{A}})$ where $\mathcal{M}^{\mathbf{H}}$ (resp. $\mathcal{M}^{\mathbf{A}}$) is the set of messages $\mathbf{H}$ (resp. $\mathbf{A}$) can send.

Definition 5.2. A *partially observable off-switch game with cheap talk* (PO-OSG-C) is a PO-OSG G along with a message system that makes the following modification to G: After both players receive their observations but before they act, each player simultaneously sends a single message from their message set.

A PO-OSG is a PO-OSG-C in which the message sets are singletons. A deterministic policy $\pi^{\mathbf{A}}$ for $\mathbf{A}$ is now a map $\Omega^{\mathbf{A}} \times \mathcal{M}^{\mathbf{H}} \to \mathcal{M}^{\mathbf{A}} \times \mathcal{A}^{\mathbf{A}}$ whose first coordinate depends only on $O^{\mathbf{A}}$, and a policy $\pi^{\mathbf{H}}$ for $\mathbf{H}$ is analogous.

5.1 Communication Cannot Decrease Payoff

Messages provide information similar to observations, so we get an analogue of Theorem 4.7 for communication: increasing the communication bandwidth between $\mathbf{H}$ and $\mathbf{A}$ cannot decrease their expected payoff in optimal policy pairs.

Definition 5.3. A message system $\mathcal{M}_1$ is *(weakly) more expressive* than $\mathcal{M}_2$ if $|\mathcal{M}_1^{\mathbf{H}}| \geq |\mathcal{M}_2^{\mathbf{H}}|$ and $|\mathcal{M}_1^{\mathbf{A}}| \geq |\mathcal{M}_2^{\mathbf{A}}|$. It is *(weakly) more expressive for* $\mathbf{H}$ if it is more expressive but $|\mathcal{M}_1^{\mathbf{A}}| = |\mathcal{M}_2^{\mathbf{A}}|$, and more expressive for $\mathbf{A}$ analogously.

Definition 5.4. A message system $\mathcal{M}_1$ is *better in optimal play* than $\mathcal{M}_2$ if, for each PO-OSG G, the expected payoff under OPPs for the PO-OSG-C $(G, \mathcal{M}_1)$ is at least the expected payoff under OPPs for the PO-OSG-C $(G, \mathcal{M}_2)$.

Theorem 5.5. *If a message system $\mathcal{M}_1$ is more expressive than $\mathcal{M}_2$, then $\mathcal{M}_1$ is better in optimal play than $\mathcal{M}_2$.*

Proof. Let G be a PO-OSG. We may assume without loss of generality that $\mathcal{M}_2^{\mathbf{H}} \subseteq \mathcal{M}_1^{\mathbf{H}}$ and $\mathcal{M}_2^{\mathbf{A}} \subseteq \mathcal{M}_1^{\mathbf{A}}$. Thus, any policy pair in $(\mathcal{M}_2, G)$, including its optimal policy pair, is a valid policy pair for $(\mathcal{M}_1, G)$. Thus the optimal expected payoff for $(\mathcal{M}_1, G)$ is at least that of $(\mathcal{M}_2, G)$. □

5.2 Unbounded Communication

Definition 5.6. We say that $\mathbf{H}$ *has unbounded communication* if $|\mathcal{M}^{\mathbf{H}}| \geq |\Omega^{\mathbf{H}}|$. We define $\mathbf{A}$ having unbounded communication analogously.

When one player has unbounded communication, additional message expressiveness cannot achieve higher payoff in optimal policy pairs. In these cases, one agent can fully communicate their observation, making that agent's observation redundant. Proposition 4.3 thus yields:

Corollary 5.7. *If $\mathbf{H}$ (resp. $\mathbf{A}$) has unbounded communication, then there is an optimal policy pair in which $\mathbf{A}$ never (resp. always) defers.*

5.3 Communication Can Have Unintuitive Effects on Shutdown Incentives

In Propositions 4.9 and 4.11 players only gained information that the other player did not already know. One might expect that expanding the message set $\mathcal{M}^{\mathbf{A}}$ makes $\mathbf{A}$ more likely to defer in optimal policy pairs, since $\mathbf{A}$ can provide $\mathbf{H}$ with information that $\mathbf{A}$ already has. However, the following proposition shows this is not the case.

Proposition 5.8. *There is a PO-OSG-C $(G, \mathcal{M})$ with the property that if we replace $\mathcal{M}$ with a message system that is more expressive for $\mathbf{A}$, then $\mathbf{A}$ plays $w(a)$ strictly less often in optimal policy pairs.*

In the same vein, we may ask if decreasing the size of $\mathcal{M}^{\mathbf{H}}$ makes $\mathbf{A}$ more likely to play $w(a)$ in optimal policy pairs. This also fails to hold.

Proposition 5.9. *There is a PO-OSG-C $(G, \mathcal{M})$ with the property that if we replace $\mathcal{M}$ with a message system that is less expressive for $\mathbf{H}$, then $\mathbf{A}$ plays $w(a)$ strictly less often in optimal policy pairs.*

Proofs for both results are given in the full paper.

6 A-Unaware Human Policies

There is a common theme in the examples above: $\mathbf{A}$ defers less often to $\mathbf{H}$ in order to better coordinate with her. Is this coordination the only source of unusual behavior? In this section, we argue that ignoring the effect of coordination cannot save us. All the unintuitive results above hold even when $\mathbf{H}$ is unaware of $\mathbf{A}$'s existence.

Moving in the opposite direction to the previous sections, we now break from the model of fully rational $\mathbf{H}$ and $\mathbf{A}$ to a model of bounded rationality. Namely, we study the most basic case of a cognitively bounded $\mathbf{H}$, in which she ignores $\mathbf{A}$'s choice of action in choosing her own.

Definition 6.1. We say $\mathbf{H}$ is $\mathbf{A}$*-unaware* if $\mathbf{H}$'s policy is given by:

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } \mathbb{E}[u_a(S) - u_o(S) \mid O^{\mathbf{H}} = o^{\mathbf{H}}] > 0, \\ \text{OFF} & \text{if } \mathbb{E}[u_a(S) - u_o(S) \mid O^{\mathbf{H}} = o^{\mathbf{H}}] < 0 \end{cases}$$

and $\mathbf{H}$ can choose arbitrarily if $\mathbb{E}[u_a(S) - u_o(S) \mid O^{\mathbf{H}} = o^{\mathbf{H}}] = 0$. If $\mathbf{H}$ is not $\mathbf{A}$-unaware, we say $\mathbf{H}$ is $\mathbf{A}$*-aware*.

Note that this expectation is *not* conditioned on $\mathbf{A}$'s action. This is the sense in which $\mathbf{H}$ is $\mathbf{A}$-unaware—$\mathbf{H}$ does not update her beliefs about the possible state based on the fact that $\mathbf{A}$ has deferred to $\mathbf{H}$. This makes coordination between $\mathbf{H}$ and $\mathbf{A}$ difficult, and means that they cannot always play an optimal policy pair. However, we can still define a notion of the *best* policy pair given that $\mathbf{H}$ is $\mathbf{A}$-unaware.

Definition 6.2. A policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ is an $\mathbf{A}$*-aware optimal policy pair* if $\pi^{\mathbf{H}}$ is the policy of an $\mathbf{A}$-aware $\mathbf{H}$ and $\pi^{\mathbf{A}}$ is a best response to $\pi^{\mathbf{H}}$.

Our motivation for studying the behavior of an $\mathbf{A}$-unaware $\mathbf{H}$ is threefold. First, it offers a more realistic model of bounded human cognition. Second, optimal policy pairs with $\mathbf{A}$-aware $\mathbf{H}$ might be computationally intractable to find. Finally, discussing an $\mathbf{A}$-unaware $\mathbf{H}$ allows us to isolate the effect of communication in PO-OSGs—an $\mathbf{A}$-unaware $\mathbf{H}$ ignores all communication from $\mathbf{A}$, even of the implicit sort considered in Section 4.5.

6.1 Making an A-Unaware H More Informed Can Decrease Payoffs

In contrast with Theorems 4.7 and 5.5, the value of information is *not* necessarily positive when $\mathbf{H}$ is $\mathbf{A}$-unaware. This is formalized in Proposition 6.3 below. Here, the notion of "better in $\mathbf{A}$-unaware optimal play" is the same as Definition 4.6 except replacing "optimal policy pairs" with "$\mathbf{A}$-unaware optimal policy pairs."

Proposition 6.3. *(a) If an observation structure $\mathcal{O}$ is more informative for $\mathbf{A}$ than $\mathcal{O}'$, then $\mathcal{O}$ is better in $\mathbf{A}$-unaware optimal play than $\mathcal{O}'$.*

(b) On the other hand, there is a PO-OSG G such that if one modifies G by making its observation structure strictly more informative for $\mathbf{H}$, then we obtain a worse expected payoff in $\mathbf{A}$-unaware optimal policy pairs.

This is shown in the full version of this paper. Proposition 6.3(b) implies that, given the choice of which observation structure to give an $\mathbf{A}$-unaware $\mathbf{H}$, $\mathbf{A}$ could have an incentive to give $\mathbf{H}$ the less informative one. This result is qualitatively similar to Emmons et al. (2024)'s examples of sensor tampering in assistance games.

6.2 Information Gain Can Have Unintuitive Effects on Shutdown Incentives When H Is A-Unaware

Other than Proposition 6.3, the results for $\mathbf{A}$-unaware $\mathbf{H}$ in $\mathbf{A}$-unaware optimal policy pairs are similar to Section 4: even when deferral *cannot* be implicit communication, making $\mathbf{H}$ more informed can cause $\mathbf{A}$ to defer less and making $\mathbf{A}$ more informed can cause it to defer more.

Proposition 6.4. *(a) There is a PO-OSG G with the property that if one modifies G by making its observation structure strictly more informative for $\mathbf{H}$, then $\mathbf{A}$ plays $w(a)$ less in $\mathbf{A}$-unaware optimal policy pairs.*

(b) There is a PO-OSG G' with the property that if one modifies G' by making its observation structure strictly less informative for $\mathbf{A}$, then $\mathbf{A}$ plays $w(a)$ less in $\mathbf{A}$-unaware optimal policy pairs.

Proof (sketch). The examples used to prove Proposition 4.9 and Proposition 4.11 can be used to prove (a) and (b) respectively: the policy pairs in the examples are optimal regardless of whether $\mathbf{H}$ is aware of $\mathbf{A}$. □

7 Discussion and Conclusion

We show that even when assuming common payoffs and human rationality, partial observability can cause AIs to avoid shutdown, and basic measures that one might expect to improve the situation can sometimes make the situation worse.

Explaining the Unintuitive Results What mechanism produces these surprising effects? To answer this question, we must carefully break down the chain that connects private information to shutdown incentives. Making either agent more informed can introduce new subsets of states in which they can choose to play the action. For instance, the additional information in Figure 3b allows the agents to take the action in every state except the -5 payoff state, but it is impossible to play the action in exactly that subset of states given the information in Figure 3a. Next, an optimal policy pair (OPP) plays the action in the optimal subset of states out of all subsets that are accessible. Policy pairs using a newly available optimal subset can involve the AI waiting more or waiting less. Figure 3 shows a case where achieving a new optimal subset requires waiting less, and in the full version of this work we present a case that requires waiting more. This chain of effects explains the unintuitive finding that providing either agent with more information is compatible with the AI waiting more or less in OPPs.

Interpreting the Formalism Why should we care that $\mathbf{A}$ sometimes does not defer to $\mathbf{H}$ in optimal policy pairs (OPPs) of PO-OSGs if these policies (by definition) maximize $\mathbf{H}$'s payoff? First, it seems helpful to understand shutdown incentives regardless of whether shutdown is good or bad. Second, if we interpret the common payoff function carefully, we find that OPPs are not always desirable. The role of the u in PO-OSGs is that the players select policies to maximize it. **If we understand u as the payoff function closest to what the human acts to maximize, this may not represent H's full preferences over outcomes.**

Most payoff function formalisms have expressivity limitations that prevent them from capturing more complex human preferences (Abel et al. 2021; Skalse and Abate 2023; Subramani et al. 2024). Therefore, maximizing payoffs may not maximize $\mathbf{H}$'s overall preferences, and avoiding shutdown to maximize payoffs may be concerning. PO-OSGs thus provide a useful framework to understand when AI assistants are incentivized to avoid shutdown, allowing designers to consider their specific deployment contexts and make tradeoffs between AI deference and payoff maximization.

Limitations and Future Work Our work focuses on optimal policy pairs and best responses, which have the advantage of applying generally to any learning algorithm that can find them. However, algorithms that fail to find these optimal solutions may exhibit behavior not captured by our results. Future work could investigate relaxing several assumptions of our analysis, notably that human feedback is free, there are common payoffs, the game runs for a single round, and the human is rational. Exploring shutdown incentives in a sequential setting seems particularly interesting, as prior work has discussed new incentives to avoid shutdown that may arise in this case (Freedman and Gleave 2022; Arbital n.d.). Another question is whether the examples we use to prove our counterintuitive results are "natural"—that is, do they arise frequently in the real world? Finally, a promising path is to explore other solution concepts in PO-OSGs, such as perfect Bayesian equilibria when $\mathbf{H}$ and $\mathbf{A}$ do not have the same prior over the state or when $\mathbf{H}$ is irrational.

Acknowledgments

The authors are grateful for the support of the Berkeley Existential Risk Initiative and Open Philanthropy's gift to the Center for Human-Compatible AI.

References

Abel, D.; Dabney, W.; Harutyunyan, A.; Ho, M. K.; Littman, M.; Precup, D.; and Singh, S. 2021. On the Expressivity of Markov Reward. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*.

Arbital. n.d. Problem of fully updated deference. Accessed: 2024-08-15.

Blackwell, D. 1951. Comparison of Experiments. In Neyman, J., ed., *Second Berkeley Symposium on Mathematical Statistics and Probability*, 93–102.

Blackwell, D. 1953. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2): 265–272.

Carey, R. 2018. Incorrigibility in the CIRL Framework. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 30–35. New York, NY, USA: Association for Computing Machinery. ISBN 9781450360128.

Carey, R.; and Everitt, T. 2023. Human Control: Definitions and Algorithms. In Evans, R. J.; and Shpitser, I., eds., *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, 271–281. PMLR.

Crawford, V. P.; and Sobel, J. 1982. Strategic Information Transmission. *Econometrica*, 50(6): 1431–1451.

Emmons, S.; Oesterheld, C.; Conitzer, V.; and Russell, S. 2024. Observation Interference in Partially Observable Assistance Games. arXiv:2412.17797.

Freedman, R.; and Gleave, A. 2022. CIRL Corrigibility is fragile. LessWrong.

Fudenberg, D.; and Tirole, J. 1991. *Game theory*. MIT Press.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2017. The off-switch game. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, 220–227. AAAI Press. ISBN 9780999241103.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative Inverse Reinforcement Learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Lang, L.; Foote, D.; Russell, S.; Dragan, A.; Jenner, E.; and Emmons, S. 2024. When Your AIs Deceive You: Challenges of Partial Observability in Reinforcement Learning from Human Feedback. arXiv:2402.17747.

Lehrer, E.; Rosenberg, D.; and Shmaya, E. 2010. Signaling and mediation in games with common interests. *Games and Economic Behavior*, 68(2): 670–682.

Marschak, J.; and Miyasawa, K. 1968. Economic Comparability of Information Systems. *International Economic Review*, 9(2): 137–174.

Omohundro, S. M. 2008. The Basic AI Drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492. NLD: IOS Press. ISBN 9781586038335.

Russell, S. 2019. *Human compatible: AI and the problem of control*. Penguin UK.

Shah, R.; Freire, P.; Alex, N.; Freedman, R.; Krasheninnikov, D.; Chan, L.; Dennis, M. D.; Abbeel, P.; Dragan, A.; and Russell, S. 2020. Benefits of Assistance over Reward Learning. In *NeurIPS Workshop on Cooperative AI*.

Skalse, J.; and Abate, A. 2023. On the limitations of Markovian rewards to express multi-objective, risk-sensitive, and modal tasks. In Evans, R. J.; and Shpitser, I., eds., *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, 1974–1984. PMLR.

Soares, N.; Fallenstein, B.; Yudkowsky, E.; and Armstrong, S. 2015. Corrigibility. In Walsh, T., ed., *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*, volume WS-15-02 of *AAAI Technical Report*. AAAI Press.

Subramani, R.; Williams, M.; Heitmann, M.; Holm, H.; Griffin, C.; and Skalse, J. 2024. On the Expressivity of Objective-Specification Formalisms in Reinforcement Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wängberg, T.; Böörs, M.; Catt, E.; Everitt, T.; and Hutter, M. 2017. A Game-Theoretic Analysis of the Off-Switch Game. In Everitt, T.; Goertzel, B.; and Potapov, A., eds., *Artificial General Intelligence*, 167–177. Cham: Springer International Publishing. ISBN 978-3-319-63703-7.