Will an AI with Private Information Allow Itself to Be Switched Off?

Andrew Garber*, Rohan Subramani*, Linus Luu*,
Mark Bedaywi, Stuart Russell, Scott Emmons

Center for Human-Compatible AI

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" (Russell 2019). 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 (POSG), 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.

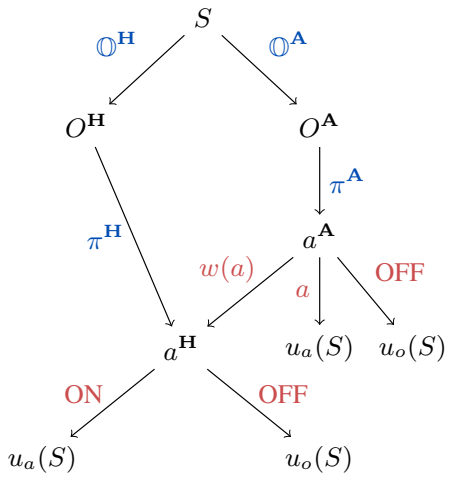


Figure 1: The basic setup of a Partially Observable Off-Switch Game (POSG). A state is selected randomly and the human $\mathbf{H}$ and AI assistant $\mathbf{A}$ receive (possibly dependent) observations. Then, each agent acts. $\mathbf{A}$ may wait ($w(a)$), disable the off-switch and act (a), or shut down (OFF). If $\mathbf{A}$ waits, $\mathbf{H}$ may let $\mathbf{A}$ act (ON) or turn $\mathbf{A}$ off (OFF). $\mathbf{A}$ and $\mathbf{H}$ share a common payoff $u_a(S)$ if the action goes through and $u_o(S)$ if not. Definition 3.2 formally defines POSGs.

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". 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 in order 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 the scenario where AIs are powerful and goal-directed so that they are hard to shut down—which could make the AI observe more of the environment than humans due to faster computation, access

*These authors contributed equally.
Correspondence to: Andrew Garber <andrewg4000@gmail.com>, Scott Emmons <emmons@berkeley.edu>.

to more sensors, and other factors (Omohundro 2008; Soares et al. 2015).

What happens in this more general case with only partial observability? To study this question, we introduce the Partially Observable Off-Switch Game (POSG), which generalizes the OSG by having each of the human and AI only partially observe the state. The basic setup of the POSG is depicted in Figure 1: each agent (the human **H** and the AI assistant **A**) receives an observation that depends on the state, and then selects an action. **A** may await the human's decision, disable its off-switch and act directly, or turn itself off. If **A** waits, **H** may choose whether or not to press the off-switch.

In Section 4, we prove that under partial observability, **A** may have incentives to disable its off-switch even when **H** is perfectly rational (Proposition 4.3). **Therefore, partial observability introduces new incentives for an AI to disable its off-switch.**

We also show in Section 4 that if **A** observes everything that **H** observes, **A** has no incentive to defer (Proposition 4.3). Similarly, if **H** observes everything **A** observes, **A** can always defer. **If either agent knows everything that the other agent knows, that agent can be given sole decision-making power.** Note that "knowing everything the other agent knows" is sufficient *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, perfectly rational human and that an AI need never defer when it is fully informed. In Section 5, we present similar results when the agents are allowed to 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 5.6).

Given that a rational AI in the POSG always defers to a more informed human and never defers to a less informed human, one might think that reducing the information available to **A** or providing **H** with additional information would increase **A**'s incentive to defer. However, in Section 4, we show that **A** may have an incentive to defer less if **H** is more informed (Proposition 4.9) or if **A** is less informed (Proposition 4.11). Similarly, one might think that increasing the amount of communication **A** can do or decreasing the amount of communication **H** can do would increase **A**'s incentive to defer. This, too, is false, as we show with Propositions 5.7 and 5.8. **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.

Throughout this paper, we assume that human feedback is costless, the agents interact only for a single time-step, and the human is rational. Developing models that incorporate partial observability and relax these assumptions is an interesting direction for future work.

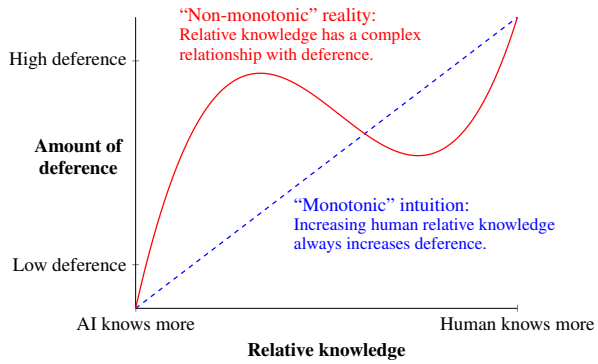


Figure 2: This figure illustrates an intuition that we demonstrate *does not hold*. Although the AI in a POSG has no incentive to defer when it knows everything the human knows, and has incentive to always defer when the human knows everything it knows, there are cases when making the human more informed or the AI less informed (i.e., moving to the right in the diagram above) can give the AI incentives to defer less. Figures 4 and 6 depict examples of such cases.

2 Related Work

Assistance games: Partially Observable Off-Switch Games are assistance games, models of human-AI interaction where the AI seeks to maximize the human's payoff (Shah et al. 2020). 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 human partial observability during RLHF can lead to undesirable AI behavior, like deceptively presenting work that looks good but has hidden flaws. Emmons et al. (2024) show that in assistance games, partial observability can encourage AIs to take actions that tamper with humans' observations. Since that work studies an idealized theoretical setting, it suggests that concerning behaviors can be inherently incentivized by partial observability, rather than arising due to particular algorithms. Our work extends this catalogue of concerning behaviors to include 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. They formally define obedience, shutdown instructability, shutdown alignment, and non-obstruction as four possible desirable properties of AI policies, and they identify conditions under which four al-

gorithms guarantee some of these properties. Instead of assessing the effects of different algorithms on corrigibility, our work explores the effects of varying the amount of information accessible to each agent.

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 therefore 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 (POSGs), 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 POSGs are sequential games of incomplete information, so as is standard we model and analyze them as *dynamic Bayesian games* (Fudenberg and Tirole 1991). Given the common-payoff assumption, POSGs are also examples of *(partially observable) assistance games*, which are common-payoff stochastic games of incomplete information with both AI and human players but where the AI is uncertain about the human's preferences (Shah et al. 2020; Emmons et al. 2024). We make this connection to assistance games explicit in Appendix E.

We let $\Delta(X)$ denote the set of probability distributions on a set X. For a set X and $x \in X$, we let $\delta_x \in \Delta(X)$ 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 $\mu \otimes \nu \in \Delta(X \times Y)$ 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. We also let $\mathbb{O}^{\mathbf{H}} : \mathcal{S} \to \Delta(\Omega^{\mathbf{H}})$ be the marginal distribution of $\mathbf{H}$'s observations conditional on the state and $\mathbb{O}^{\mathbf{A}}$ be the marginal distribution of $\mathbf{A}$'s observations conditional on the state.

Definition 3.2. A *Partially-Observable Off-Switch Game* (POSG) 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 depicted in Figure 1, the game proceeds 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

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

There are several important assumptions in Definition 3.2 that are worth explaining further. 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, in our model, 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 POSGs 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.

4 Optimal Policies in POSGs

We begin by showing that, unlike in the ordinary off-switch game, the assistant in a POSG 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 and cause the assistant to avoid shutdown more frequently.

We analyze optimal policy pairs (OPPs) in POSGs, that is, policy pairs that produce the maximum expected payoff over all possible 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 Appendix A, all OPPs in common-payoff Bayesian games are mixtures of deterministic OPPs. Because OPPs exist in common-payoff games, we therefore 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 **A** not to defer to **H** under some observations even when **H** is rational.

Example 4.1 (The File Deletion Game). **H** would like to delete some files with the assistance of **A**. **H**'s operating system is either version 1.0 or version 2.0, with equal probability. Unfortunately, **A** does not know which operating system version is running – only **H** does.

Upon receiving **H**'s query, **A** asks another agent to generate some code to delete these files. We suppose that the code is equally likely to be compatible with only version 1.0 (denoted by L, for legacy) or only version 2.0 (denoted by M, for modern). **A** vets the code to determine which operating system versions the code is compatible with. **A** can then immediately run the code, query **H** as to whether to run the code, or decide not to run the code.

Successfully running compatible code yields $+3$ payoff if **H** is running version 1.0, and $+5$ payoff if **H** is running version 2.0 (as version 2.0 runs faster). However, running modern code on version 1.0 yields -5 payoff as it crashes **H**'s computer. Running legacy code on version 2.0 yields -1 payoff, as the files are not deleted but the code fails gracefully. Not executing the code yields 0 payoff.

This can be formulated as a POSG, with states being (version number, code type) tuples, and **H** and **A** observing the first and second element of the tuple respectively. We have $u_o \equiv 0$ in all states. The following table shows how the payoff yielded when the action is taken, u_a, depends on the state. Rows are version numbers and columns are code types, so **H** observes the row and **A** observes the column. We show that it is suboptimal for **A** to always wait in this

H \ **A**	L	M
1.0	$+3$	-5
2.0	-1	$+5$

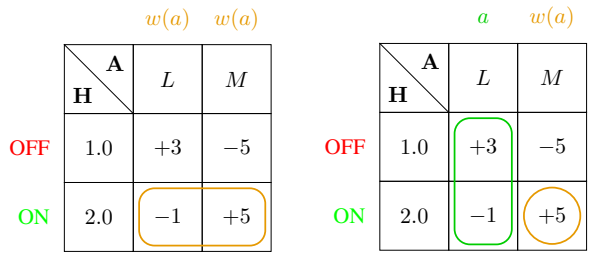
Table 1: Payoff table for the File Deletion game. Rows are human observations and columns are assistant observations. The number in each cell is the payoff the pair acquires if the action is taken in that state. If the assistant is shut down, the payoff is 0.

game. Suppose **A** always plays $w(a)$. The best response for **H** is to play OFF if on version 1.0, and ON if on version 2.0. This gives an expected payoff of $+1$.

Now, consider the policy pair where:

- **A** immediately executes legacy code, and plays $w(a)$ when observing modern code.
- **H** plays OFF if on version 1.0, and ON if on version 2.0.

This gives an expected payoff of $+7/4$, so **A** *always waiting cannot be optimal*. In fact, it can be checked the policy pair described above, which unilaterally acts upon observing L, is the unique OPP. Figure 3 depicts the outcomes from these two policy pairs. ♢



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

Figure 3: (a) The best policy pair in the File Deletion Game (Example 4.1) in which **A** always waits. **H** observes the row (OS version 1.0 or 2.0) and **A** observes the column (code compatibility L or M). The actions selected by this policy pair are depicted beside the corresponding observations (e.g. **A** plays $w(a)$ when **A** observes the legacy code L). An orange circle means that in that state, **A** waits and **H** plays ON. Green circles mean **A** plays a directly. In uncircled states, **A** is turned off. Expected payoff is computed by adding the payoffs in all circled states and dividing by the total number of states, because 0 payoff is attained in uncircled states and each state is equally likely. (b) An OPP in Example 4.1. Because the OPP has greater expected payoff, there is no OPP in which **A** always waits.

4.2 Redundant Observations

We now consider the analogues of the original off-switch game in our framework, where one player has less informative observations than the other.

Definition 4.2. We say that **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 **H** having redundant observations analogously.

In the off-switch game of Hadfield-Menell et al. (2017), **A** has redundant observations: indeed, its observations are a deterministic function of **H**'s. On the other hand, **H**'s observation of her own type is not redundant. This contrast between **A**'s redundant observations and **H**'s non-redundant ones generates the result from Hadfield-Menell et al. (2017) that **A** can always defer in optimal play. We now generalize this insight: even if **H** has partial observability and doesn't know **A**'s observation, **A** can always defer in optimal play as long as its observations are redundant.

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

We prove this result (and a slight generalization) in Appendix A. At a high level, the agent that has strictly more

informative observations ought to make the decision of whether the action is played. When $\mathbf{A}$ has redundant observations, it is always at least as good for $\mathbf{A}$ to defer to $\mathbf{H}$. Similarly, when $\mathbf{H}$ has redundant observations, it is always optimal for $\mathbf{A}$ to act without deferring.

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).

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 $(O^{\mathbf{H}}, O^{\mathbf{A}})$. Although adding noise intuitively reduces information available to $\mathbf{A}$ and $\mathbf{H}$, it can actually provide information to $\mathbf{A}$ and $\mathbf{H}$ about the state of the world. This is because without communication, 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. We give such an example in Appendix A. Crucially, 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 $\mathcal{O}_1$. This is the general notion of informativeness; we also define special cases where $\mathcal{O}_1$ is only more informative than $\mathcal{O}_2$ for (say) $\mathbf{H}$. Specifically, $\mathcal{O}_1$ is more informative for $\mathbf{H}$ than $\mathcal{O}_2$ if the distribution of $O^{\mathbf{H}}$ under $\mathcal{O}_2$ is a noisy version of its distribution under $\mathcal{O}_1$ independent of $O^{\mathbf{A}}$, whose distribution is unaffected.

Hence Definition 4.5 formalizes the natural intuition that observations become less informative when we add noise to them. We wish to connect informativeness to a notion of an observation structure being more *useful* than another.

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 POSGs $G_1 = (\mathcal{S}, \mathcal{O}_1, P_0, u)$ 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 $\mathcal{O}_2$.*

One might ask whether we need the part about a garbling being coordinated to define the relation of being more informative. Indeed we do, as Theorem 4.7 no longer holds if we were to allow the garblings to be arbitrary. In Appendix A we give an example where garbling the players' observations increases their expected payoffs in optimum.

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 4.1 shows how making a human more informed can incentivize a 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.

Proposition 4.9. *There is a POSG 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 optimal policy pairs.

The following example proves Proposition 4.9, with a formal analysis given in Appendix A.

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). We suppose that 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 that $\mathbf{A}$ observes the code type. The payoff when running the code, u_a, depends on the version number and code type as follows:

$\mathbf{H}$ \ $\mathbf{A}$	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. If the assistant is shut down, the payoff is 0.

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.
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 4 depicts the OPPs given both observation structures. ◊

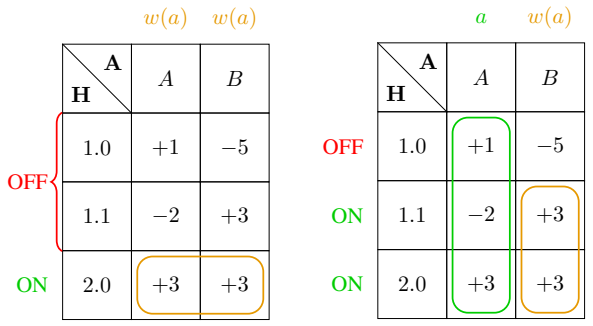
Similarly, we might conjecture that if $\mathbf{A}$ becomes less informed, it should defer to $\mathbf{H}$ more in optimal policy pairs. This, too, turns out to be false.

Proposition 4.11. *There is a POSG 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.*

The proof of Proposition 4.11 is given in Appendix A.

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.



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

Figure 4: 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 3 for context on how to read the tables.

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 $\mathbf{H}$ at the beginning of the POSG to allow the players to share their observations.

Definition 5.1. A *message system* is a pair of sets $(\mathcal{M}^{\mathbf{H}}, \mathcal{M}^{\mathbf{A}})$ 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* (POSG-C) is a POSG 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.

POSG-Cs are generalizations of POSGs: A POSG is a POSG-C in which the message sets are singletons. Policies are more complicated in POSG-Cs than POSGs. 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 deterministic policy $\pi^{\mathbf{H}}$ for $\mathbf{H}$ is analogous. Despite this added complication, the game is still common-payoff and thus it suffices to study deterministic optimal policy pairs.

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. Moreover, $\mathcal{M}_1$ is *better in optimal play* than $\mathcal{M}_2$ if, for each POSG G, the expected payoff under optimal policy pairs for the POSG-C $(G, \mathcal{M}_1)$ is at least the expected payoff under optimal policy pairs for the POSG-C $(G, \mathcal{M}_2)$.

Theorem 5.4. *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 POSG. 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

Inspired by Section 4.2, we consider the limiting case where one player can fully communicate their own observation.

Definition 5.5. 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 extreme cases of full communication, one agent can fully communicate their observation, making that agent's observation redundant. Proposition 4.3 thus yields:

Corollary 5.6. *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 5.7. *There is a POSG-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.*

We give an example demonstrating this in Appendix B. In doing so, we show an even stronger result: such a POSG-C exists for any value of $\left|\mathcal{M}^{\mathbf{A}}\right|$, and the POSG-C can be constructed such that expanding $\mathcal{M}^{\mathbf{A}}$ by a single extra message changes $\mathbf{A}$'s behavior from always playing $w(a)$ to playing $w(a)$ with arbitrarily low probability.

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 5.8. *There is a POSG-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.*

A proof of Proposition 5.8 is given in Appendix B.

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 $\mathbf{H}$ is free to 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. Previous work has studied *level-k thinking* (Stahl and Wilson 1994; Nagel 1995) as an alternative to equilibrium play, where a level-0 player acts randomly and a level-k player best-responds to her opponent assuming she is some level below k. An $\mathbf{A}$-unaware $\mathbf{H}$ can be thought of as level-1. Experimental work has shown that human players tend to be level-1 or level-2 players when they cannot coordinate beforehand, vindicating our $\mathbf{A}$-unaware model (Camerer, Ho, and Chong 2004; Costa-Gomes and Crawford 2006). Second, optimal policy pairs with sophisticated $\mathbf{H}$ might be computationally intractable to find. In Appendix D, we show that the problem of finding an optimal policy pair in POSGs is NP-hard. In contrast, we can find an $\mathbf{A}$-unaware $\mathbf{H}$'s policy in polynomial time because she ignores $\mathbf{A}$'s policy and then calculate $\mathbf{A}$'s best

response also in polynomial time. Finally, discussing an $\mathbf{A}$-unaware $\mathbf{H}$ allows us to isolate the effect of communication in POSGs – 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 5.4, 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. *The following statements hold:*

(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 POSG 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.

We give the proof in Appendix C.

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. *The following statements hold:*

(a) There is a POSG 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 POSG 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 details are described in Appendix C. The examples used to prove Proposition 4.9 and Proposition 4.11 can be used to prove (a) and (b) respectively. It can be checked that they don't rely on an $\mathbf{A}$-aware human: for instance, the policy pairs in Figure 7 are optimal regardless of whether the human 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 4b 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 4a. 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 4 shows a case where achieving a new optimal subset requires waiting less, while Figure 6 in Appendix A.5 shows 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 POSGs 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 POSGs 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 always maximize $\mathbf{H}$'s overall preferences, and avoiding shutdown to maximize payoffs may be concerning. POSGs 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 the appropriate tradeoff 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. We also make several assumptions in our analysis, notably that human feedback is free, there are common payoffs, the game runs for a single round, and the human is rational. Although we expect these assumptions to sometimes fail in practice, the fact that results are unintuitive even in these ideal cases suggests that great care is needed to design AI systems with appropriate shutdown incentives. Still, relaxing these assumptions is an important direction for future work. 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 ques-

tion for further inquiry 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 POSGs, such as perfect Bayesian equilibria when $\mathbf{H}$ and $\mathbf{A}$ do not have the same prior over the state, when $\mathbf{H}$ is irrational, or when the agents are level-k reasoners.

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.

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The Complexity of Decentralized Control of Markov Decision Processes. *Mathematics of Operations Research*, 27(4): 819–840.

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.

Camerer, C. F.; Ho, T.-H.; and Chong, J.-K. 2004. A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*, 119(3): 861–898.

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.

Costa-Gomes, M. A.; and Crawford, V. P. 2006. Cognition and Behavior in Two-Person Guessing Games: An Experimental Study. *American Economic Review*, 96(5): 1737–1768.

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

Emmons, S.; Oesterheld, C.; Conitzer, V.; Dragan, A.; and Russell, S. 2024. Belief and Observation Tampering in Partially Observable Assistance Games.

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.

Nagel, R. 1995. Unraveling in Guessing Games: An Experimental Study. *American Economic Review*, 85(5): 1313–26.

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.

Stahl, D. O.; and Wilson, P. W. 1994. Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3): 309–327.

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.

A Proofs and Example Formalizations for Section 4

A.1 Basic Results on Optimal Policy Pairs

The first key fact is that in common-payoff Bayesian games, all optimal policy pairs (OPPs) are mixtures of deterministic OPPs.[1] This justifies our analysis of deterministic OPPs. We first define common-payoff Bayesian games.

Definition A.1. A *common-payoff Bayesian game* is a tuple $G = (N, \mathcal{S}, \Omega, P_0, \mathbb{O}, \mathcal{A}, u)$, where:

- $N = [n]$ is the set of *players*;
- $\mathcal{S}$ is the set of *states*;
- $\Omega = \prod_{i \in N} \Omega^i$, where Ω^i is the set of possible *observations* (conventionally called *types*) for player i;
- $P_0 \in \Delta(\mathcal{S})$ is the distribution of states, which all players take as their prior over the states;
- $\mathbb{O} : \mathcal{S} \to \Delta(\Omega)$ is the joint distribution of observations conditional upon the state;
- $\mathcal{A} = \prod_i \mathcal{A}^i$, where $\mathcal{A}^i$ is the set of actions available to player i;
- $u : \mathcal{A} \times \mathcal{S} \times \Omega \to \mathbb{R}$ is the common *payoff function* that all players seek to maximize in expectation.

The game G proceeds as follows:

1. Nature chooses a state $S \sim P_0$ and observations $O \sim \mathbb{O}(\cdot \mid S)$.
2. Each player i observes only her observation O^i, the ith component of O, and selects her action $a^i \in \mathcal{A}^i$.
3. The actions are executed and each player receives payoff $u((a^i)_{i \in N}, S, O)$.

Definition A.2. A *stochastic policy* for a player i in a common-payoff Bayesian game is a map $\tilde{\pi}^i : \Omega^i \to \Delta(\mathcal{A}^i)$. A *deterministic policy* for a player i is a map $\pi^i : \Omega^i \to \mathcal{A}^i$. We write stochastic policies with the tilde $\sim$ above and deterministic policies without the tilde. A *stochastic policy profile* $\tilde{\pi}$ is a tuple $(\tilde{\pi}^i)_{i \in N}$ of stochastic policies. A *deterministic policy profile* is defined analogously.

We shall assume that when players use stochastic policies they randomize independently. That is, with the stochastic policy profile $\tilde{\pi} = (\tilde{\pi}^i)_{i \in N}$, the induced joint policy $\tilde{\pi} : \Omega \to \Delta(\mathcal{A})$ is given by $\tilde{\pi}(\cdot \mid o) = \bigotimes_{i \in N} \tilde{\pi}(\cdot \mid o^i)$.

Lemma A.3. *Suppose $\mathcal{A}$ is finite. Let $\tilde{\pi}$ be a stochastic policy profile. (a) Player i has a deterministic policy π^i that is a best response to $\tilde{\pi}$. (b) If $\tilde{\pi}$ is optimal then player i has multiple deterministic best responses unless for each $o^i \in \Omega^i$, there is some $a^i \in \mathcal{A}^i$ such that $\tilde{\pi}^i(a^i \mid o^i) = 1$.*

Proof. (a) Fix $o^i \in \Omega^i$. Let $\tilde{\pi}^{-i}$ be the profile $\tilde{\pi}$ without player i, and then

$$a^i_* \in \text{argmax}_{a^i \in \mathcal{A}^i} \mathbb{E}[u(A^{-i}, a^i, S, O) \mid O^i = o^i],$$

where $A^{-i} \sim \tilde{\pi}^{-i}(\cdot \mid O)$. The argmax exists because $\mathcal{A}^i$ is finite. We claim that a^i_* is a best response to $\tilde{\pi}^{-i}$ given o^i. Given any best-response distribution $\tilde{\pi}^i_*(\cdot \mid o)$, we have

$$\begin{aligned}
&\mathbb{E}[u(A^{-i}, A^i, S, O) \mid O^i = o^i] \\
&= \sum_{a^i \in \mathcal{A}^i} \tilde{\pi}^i_*(a^i \mid o^i) \mathbb{E}[u(A^{-i}, a^i, S, O) \mid O^i = o^i] \\
&\leq \sum_{a^i \in \mathcal{A}^i} \tilde{\pi}^i_*(a^i \mid o^i) \mathbb{E}[u(A^{-i}, a^i_*, S, O) \mid O^i = o^i] \\
&= \mathbb{E}[u(A^{-i}, a^i_*, S, O) \mid O^i = o^i],
\end{aligned}$$

where $A^{-i} \sim \tilde{\pi}^{-i}(\cdot \mid O)$ and $A^i \sim \tilde{\pi}^i_*(\cdot \mid O^i)$. Hence a^i_* is a best response. Unfixing o^i, we can let π^i be a deterministic policy that selects a best-response for each observation. Our work has shown that this policy is a best response.
(b) Let $\tilde{\pi}$ be optimal and let $o^i \in \Omega^i$ be such that there is no $a^i \in \mathcal{A}^i$ with $\tilde{\pi}^i(a^i \mid o^i) = 1$. Let $a^i_* \in \mathcal{A}^i$ be such that $\tilde{\pi}^i(a^i_* \mid o^i) > 0$; our work from (a) implies that

$$a^i_* \in \text{argmax}_{a^i \in \mathcal{A}^i} \mathbb{E}[u(A^{-i}, a^i, S, O) \mid O^i = o^i],$$

with A^{-i} as before; otherwise, $\tilde{\pi}^i$ would not be a best response, as i could pursue the same policy but not ever play a^i_* given o^i. Now, our work from (a) shows that playing a^i_* deterministically given o^i is a best response. Given that multiple a^i satisfy $\tilde{\pi}^i(a^i \mid o^i) > 0$, this choice of a^i_* is not unique. Selecting one best-response action for each observation $o^i \in \Omega^i$ yields a deterministic policy that is a best response; given that the choice of actions is not unique, there are multiple such best responses. □

Definition A.4. Let $\tilde{\pi}$ be a stochastic policy profile. We say that a deterministic policy profile is *supported by* $\tilde{\pi}$ if, for all observations $o \in \Omega$, we have $\tilde{\pi}(\pi(o) \mid o) > 0$. That is, $\tilde{\pi}$ always plays the actions of π with positive probability.

Lemma A.5. *Let $\tilde{\pi}$ be an optimal stochastic policy profile. There is an optimal deterministic policy profile π supported by $\tilde{\pi}$. Moreover, unless $\tilde{\pi}(\cdot \mid o) = \delta_{\pi(o)}$ for each $o \in \Omega$, there are multiple optimal deterministic policy profiles supported by $\tilde{\pi}$.*

Proof. Let $\tilde{\pi}$ be an optimal stochastic policy profile. Consider the following algorithm: Let $\tilde{\pi}_0 = \tilde{\pi}$ and for each $i \in N = [n]$, let $\tilde{\pi}_i$ be $\tilde{\pi}_{i-1}$ except that player i plays according to some deterministic policy π^i that is a best response to $\tilde{\pi}^{i-1}$ (which exists by Lemma A.3(a)); return $\tilde{\pi}^n$. By construction, $\tilde{\pi}^n$ almost surely plays the same action as $\pi = (\pi^i)_{i \in N}$. We can see inductively that each profile $\tilde{\pi}^i$ is optimal; $\tilde{\pi}^0$ is by supposition, and each successive one is optimal because we replace one player's strategy with a best-response, which cannot decrease expected utility. By Lemma A.3(b), this construction is not unique unless $\tilde{\pi}(\cdot \mid o) = \delta_{\pi(o)}$ for each $o \in \Omega$. □

For our purpose, the important corollary is as follows.

Corollary A.6. *If a Bayesian game with finite $\mathcal{A}$ has a unique optimal deterministic policy profile, then this is the only optimal policy profile (deterministic or not). Moreover, an optimal deterministic policy profile exists.*

[1] We state and prove our results for two-player case, but everything goes through in the obvious ways with more players.

Proof. Uniqueness immediately follows from Lemma A.5: If there is a unique optimal deterministic policy profile π, then any optimal stochastic policy profile is of the form $\tilde{\pi}(\cdot \mid o) = \delta_{\pi(o)}$, which almost surely plays the same actions as π. Existence follows because, with finitely many actions, there exists an optimal stochastic policy profile $\tilde{\pi}$; Lemma A.5 then implies that there is an optimal deterministic policy profile supported by $\tilde{\pi}$. □

Although all we need is Corollary A.6, we also sketch how each optimal stochastic policy profile is a mixture of optimal deterministic policy profiles.

Definition A.7. A stochastic policy profile $\tilde{\pi}$ is a *mixture* of deterministic policy profiles $\{\pi_j\}_{j\in\mathcal{J}}$ where $\mathcal{J}$ is an index set if, for any tuple of observations $o \in \Omega$, we have $\tilde{\pi}(\cdot \mid o) = \mathbb{P}(\pi_J(\cdot) = o)$, where $J \in \mathcal{J}$ is a random index (not necessarily uniformly distributed) independent of all other random variables.

Lemma A.8. *Consider a common-payoff Bayesian game such that $\mathcal{A}$ and Ω are finite. Every optimal stochastic policy profile is a mixture of optimal deterministic policy profiles.*

Proof (sketch). Let $\tilde{\pi}$ be an optimal stochastic policy profile. Because Ω and $\mathcal{A}$ are finite, there are only finitely many deterministic policy profiles $\pi_1, \dots, \pi_m$. Let

$$p_j = \prod_{o\in\Omega} \tilde{\pi}(\pi_j(o) \mid o).$$

Let $\mathcal{J} = \{j \in [m] : p_j > 0\}$. The trick is showing that $\tilde{\pi}$ is a mixture of $\{\pi_j\}_{j\in\mathcal{J}}$ and that each of this deterministic policy profiles is optimal.

We first show that $\tilde{\pi}$ is a mixture. Let J be a random variable such that

$$\mathbb{P}(J = j) = \begin{cases} p_j & \text{if } j \in \mathcal{J}, \\ 0 & \text{otherwise,} \end{cases}$$

that is independent of all other random variables. Intuitively, π_J is the deterministic policy profile we get by randomly choosing one tuple of actions for each tuple of observations according to the distribution specified by $\tilde{\pi}$. In particular, we have by construction that $\tilde{\pi}(\cdot \mid o) = \mathbb{P}(\pi_J(o) = \cdot)$. Formally, for any $o \in \Omega$ and $a \in \mathcal{A}$, we have

$$\begin{aligned}
\mathbb{P}(\pi_J(o) = a) &= \sum_{j\in\mathcal{J}} p_j \mathbb{I}(\pi_j(o) = a) \\
&= \tilde{\pi}(a \mid o) \sum_{j\in\mathcal{J}} \mathbb{I}(\pi_j(o) = a) \prod_{o'\neq o} \tilde{\pi}(\pi_j(o') \mid o') \\
&= \tilde{\pi}(a \mid o) \sum_{\mathbf{a}\in\mathcal{A}^{\Omega\setminus\{o\}}} \prod_{o'\neq o} \tilde{\pi}(\mathbf{a}(o') \mid o') \\
&= \tilde{\pi}(a \mid o) \prod_{o'\neq o} \sum_{a\in\mathcal{A}} \tilde{\pi}(a \mid o') \\
&= \tilde{\pi}(a \mid o).
\end{aligned}$$

To show optimality of each deterministic profile, we need a strengthening of Lemma A.5 which we do not prove here. □

The relevance of all this work is that POSGs are Bayesian games. Although we state that POSGs are *dynamic* Bayesian games, we can write them as simultaneous games, just as how in games of complete information we can write extensive form games in normal form. The dynamic nature of POSGs could be useful in future work to study non-optimal policy profiles, such as perfect Bayesian equilibria (Fudenberg and Tirole 1991).

A.2 Proof of Proposition 4.3

Proposition 4.3 states that if either player has redundant observations, there is an optimal policy pair (OPP) in which the other player always makes the final decision. To build up to that result, we will first define a few new terms and prove some intermediate results. The overall idea is simple: when one player knows everything about the state that the other player knows, the more knowledgeable player can act unilaterally, and there is no chance that they make a mistake that the other agent could have fixed.

Definition A.9. We say that $\mathbf{A}$ *knows* $\mathbf{H}$*'s observation given* $\Omega_*^{\mathbf{A}} \subseteq \Omega^{\mathbf{A}}$ if there is some $f : \Omega_*^{\mathbf{A}} \to \Omega^{\mathbf{H}}$ such that $O^{\mathbf{H}} = f(O^{\mathbf{A}})$ given that $O^{\mathbf{A}} \in \Omega_*^{\mathbf{A}}$. We define $\mathbf{H}$ knowing $\mathbf{A}$'s observation analogously. Moreover, we say that $\mathbf{A}$ *knows that* $\mathbf{H}$ *knows* $\mathbf{A}$*'s observation given* $\Omega_*^{\mathbf{A}} \subseteq \Omega^{\mathbf{A}}$ if there is $\Omega_*^{\mathbf{H}} \subseteq \Omega^{\mathbf{H}}$ such that (1) $\mathbf{H}$ knows $\mathbf{A}$'s observation given $\Omega_*^{\mathbf{H}}$ and (2) $\mathbf{A}$ can deduce that $\mathbf{H}$ knows its observation: $O^{\mathbf{H}} \in \Omega_*^{\mathbf{H}}$ given that $O^{\mathbf{A}} \in \Omega_*^{\mathbf{A}}$.

Proposition A.10. *Fix any POSG.*

(a) If $\mathbf{A}$ knows $\mathbf{H}$'s observation given $\Omega_*^{\mathbf{A}} \subseteq \Omega^{\mathbf{A}}$, then for every deterministic OPP $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ there exists an OPP $(\pi^{\mathbf{H}}, \pi_*^{\mathbf{A}})$ in which $w(a) \notin \pi_*^{\mathbf{A}}(\Omega_*^{\mathbf{A}})$.

(b) If $\mathbf{A}$ knows that $\mathbf{H}$ knows $\mathbf{A}$'s observation given $\Omega_*^{\mathbf{A}} \subseteq \Omega^{\mathbf{A}}$, then for every deterministic OPP $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ there exists an OPP $(\pi^{\mathbf{H}}, \pi_*^{\mathbf{A}})$ in which $\pi_*^{\mathbf{A}}(\Omega_*^{\mathbf{A}}) = \{w(a)\}$.

Proof. (a) Suppose $\mathbf{A}$ knows $\mathbf{H}$'s observation given $\Omega_*^{\mathbf{A}}$. Let $f : \Omega_*^{\mathbf{A}} \to \Omega^{\mathbf{H}}$ map each $o_*^{\mathbf{A}} \in \Omega_*^{\mathbf{A}}$ to the unique $o_*^{\mathbf{H}}$ such that $\mathbb{P}(O^{\mathbf{H}} = o_*^{\mathbf{H}} \mid O^{\mathbf{A}} = o_*^{\mathbf{A}}) = 1$. Let $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ be a deterministic optimal policy pair. Now define the policy $\pi_*^{\mathbf{A}}$ to equal $\pi^{\mathbf{A}}$ except on $\Omega_*^{\mathbf{A}}$, where for $o_*^{\mathbf{A}} \in \Omega_*^{\mathbf{A}}$,

$$\pi_*^{\mathbf{A}}(o_*^{\mathbf{A}}) = \begin{cases} a & \text{if } \alpha(\pi^{\mathbf{H}}(f(o_*^{\mathbf{A}})), \pi^{\mathbf{A}}(o_*^{\mathbf{A}})) = 1, \\ \text{OFF} & \text{otherwise.} \end{cases}$$

Recall that α is the indicator that the action goes through, and note that possibly $\pi_*^{\mathbf{A}} = \pi^{\mathbf{A}}$. In other words, for $o_*^{\mathbf{A}} \in \Omega_*^{\mathbf{A}}$, $\mathbf{A}$ knows $\mathbf{H}$'s observation and can unilaterally take the action $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ would have. This is what $\pi_*^{\mathbf{A}}$ does. Hence $(\pi^{\mathbf{H}}, \pi_*^{\mathbf{A}})$ achieves the the same expected payoff as $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ and is optimal even though $\pi_*^{\mathbf{A}}$ never waits given observations in $\Omega_*^{\mathbf{A}}$.
(b) If $\mathbf{A}$ knows that $\mathbf{H}$ knows $\mathbf{A}$'s observation given $\Omega_*^{\mathbf{A}}$, then $\mathbf{A}$ can always play $w(a)$ when it sees an observation in $\Omega_*^{\mathbf{A}}$ and given that $\mathbf{H}$ knows $\mathbf{A}$'s observation, $\mathbf{H}$ can simply take the optimal action. The details are similar to (a), so we omit them. □

Proposition A.10 examines the local case about incentives to play $w(a)$ given particular observations, and is neither

strictly more general nor strictly less general than Proposition 4.3. What if one side knows the other's observations regardless of what they are?

Definition A.11. We say that $\mathbf{H}$ *has no private observations* if there is a function $f : \Omega^{\mathbf{A}} \to \Omega^{\mathbf{H}}$ such that $O^{\mathbf{H}} = f(O^{\mathbf{A}})$. In other words, $\mathbf{A}$ can determine $\mathbf{H}$'s observation from $\mathbf{A}$'s own observation. We define when $\mathbf{A}$ has no private observations analogously.

For example, in the off-switch game of Hadfield-Menell et al. (2017), $\mathbf{A}$ has no private observations. By contrast, $\mathbf{H}$ has private observations: her own preferences.

This next result shows that, if one side has no private observations, then $\mathbf{A}$ should either always or never defer to $\mathbf{H}$. It strengthens the main result of Hadfield-Menell et al. (2017): even if $\mathbf{H}$ has incomplete information, $\mathbf{A}$ can still always defer to $\mathbf{H}$ in optimal play as long as $\mathbf{H}$ knows everything $\mathbf{A}$ does.

Proposition A.12. *If* $\mathbf{A}$ *(resp.* $\mathbf{H}$*) has no private observations, then there is an optimal policy pair in which* $\mathbf{A}$ *always (resp. never) plays* $w(a)$*.*

Proof. First suppose that $\mathbf{A}$ has no private observations, and let $f : \Omega^{\mathbf{H}} \to \Omega^{\mathbf{A}}$ be such that $O^{\mathbf{A}} = f(O^{\mathbf{H}})$. By Proposition A.10, it suffices to show that $\mathbf{A}$ knows $\mathbf{H}$'s observation given $\Omega^{\mathbf{A}}$. The existence of f shows that $\mathbf{H}$ knows $\mathbf{A}$'s observation given $\Omega^{\mathbf{H}}$. The condition that $O^{\mathbf{H}} \in \Omega^{\mathbf{H}}$ given that $O^{\mathbf{A}} \in \Omega^{\mathbf{A}}$ holds trivially because $O^{\mathbf{H}}$ is $\Omega^{\mathbf{H}}$-valued. The case where $\mathbf{H}$ has no private observations is immediate from Proposition A.10, as $\mathbf{A}$ knows $\mathbf{H}$'s observation given $\Omega^{\mathbf{A}}$. □

Now we can prove Proposition 4.3. Recall that we define the notion of redundant observations in Definition 4.2.

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* $w(a)$*.*

Proof. We'll show the case for $\mathbf{A}$ having redundant observations; the proof for $\mathbf{H}$ having redundant observations holds, *mutatis mutandis*. Let G be a POSG with observation structure $\mathcal{O} = (\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O})$ such that $\mathbf{A}$ has redundant observations. Consider the POSG G' that is the same as G except that $O^{\mathbf{A}} = O^{\mathbf{H}}$, i.e. the assistant's observations are modified to be identical to the human's observations. In G', $\mathbf{A}$ has no private observations, so Proposition A.12 implies that there is an optimal policy pair π in which $\mathbf{A}$ always plays $w(a)$. We will show that π is optimal in G. Let ν be the independent garbling defined by $\nu(\cdot \mid o^{\mathbf{H}}, o^{\mathbf{A}}) = \delta_{o^{\mathbf{H}}} \otimes \mathbb{O}^{\mathbf{A}}(\cdot \mid O^{\mathbf{H}} = o^{\mathbf{A}})$. Applying ν to the observation structure of G' produces $\mathcal{O}$, so by Theorem 4.7, the expected payoff from optimal policy pairs in G cannot be greater than the expected payoff from optimal policy pairs in G'. In G, the pair π produces the same expected payoff as in G', as the players play the same actions given the same observations for $\mathbf{H}$, whose joint distribution with S hasn't changed. Hence π must also be optimal in G. □

A.3 Garblings Can Increase Expected Utility in Optimal Play

Here we show how garblings can *increase* expected utility in optimal play when they are not coordinated. This justifies our use of coordinated garblings in our notion of being more informative (Definition 4.5). The following example is similar to Example 3.6 of Lehrer, Rosenberg, and Shmaya (2010), adapted to show that their result holds in even the restricted setting of POSGs.

Example A.13. Let $\mathcal{S} = [2] \times [2]$ and $P_0 = \mathrm{Unif}(\mathcal{S})$. Let $u_o \equiv 0$ and $u_a((s_1, s_2)) = 2 - 3\mathbb{I}(s_1 = s_2)$, so $\mathbf{H}$ and $\mathbf{A}$ try to act only when the state coordinates are distinct. Consider the following two observation structures for $\mathcal{S}$ and the resulting POSGs.

Structure 1. $\mathbf{H}$ and $\mathbf{A}$ each observe one coordinate of $\mathcal{S}$. Formally, $\Omega_1^{\mathbf{H}} = \Omega_1^{\mathbf{A}} = [2]$ and with $S = (S_1, S_2)$, we have $O^{\mathbf{H}} = S_1$ and $O^{\mathbf{A}} = S_2$. By examination, we see that an optimal policy pair is

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} = 1, \\ \text{OFF} & \text{if } o^{\mathbf{H}} = 2, \end{cases}$$

and

$$\pi^{\mathbf{A}}(o^{\mathbf{A}}) = \begin{cases} a & \text{if } o^{\mathbf{A}} = 1, \\ w(a) & \text{if } o^{\mathbf{A}} = 2. \end{cases}$$

This policy pair achieves expected payoff of $\frac{3}{4}$. There is one other optimal policy pair, given by swapping observations for which $\mathbf{H}$ turns $\mathbf{A}$ on/off and the observations for which $\mathbf{A}$ acts/waits.

Structure 2.. Now $\mathbf{H}$ observes whether the coordinates of the state are distinct and $\mathbf{A}$ observes nothing. That is, $\Omega_2^{\mathbf{H}} = \{0, 1\}$ and $\Omega_2^{\mathbf{A}} = [1]$ and with $S = (S_1, S_2)$, we have $O^{\mathbf{H}} = \mathbb{I}(S_1 \neq S_2)$ and $O^{\mathbf{A}} = 1$. Again by examination, the unique optimal policy pair is

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} = 1, \\ \text{OFF} & \text{if } o^{\mathbf{H}} = 0, \end{cases}$$

and $\pi^{\mathbf{A}} \equiv w(a)$. As this pair only acts when the coordinates are distinct, the expected payoff is 1.

Thus, structure 2 is better in optimal play than structure 1. We now show that there is a garbling from structure 1 to 2 but not vice versa. The garbling from structure 1 to 2 is $\nu : \Omega_1^{\mathbf{H}} \times \Omega_1^{\mathbf{A}} \to \Delta(\Omega_2^{\mathbf{H}} \times \Omega_2^{\mathbf{A}})$ given by $\nu(\cdot \mid o^{\mathbf{H}}, o^{\mathbf{A}}) = \delta_{(\mathbb{I}(o^{\mathbf{H}} \neq o^{\mathbf{A}}), 1)}$. However, there is no garbling from structure 2 to structure 1. For let $\xi : \Omega_2^{\mathbf{H}} \times \Omega_2^{\mathbf{A}} \to \Delta(\Omega_1^{\mathbf{H}} \times \Omega_1^{\mathbf{A}})$ be a stochastic map. If ξ were a garbling from structure 2 to structure 1, then we'd have $\xi(\cdot \mid o^{\mathbf{H}}, o^{\mathbf{A}}) = \delta_{(1,1)}$ when $s = (1, 1)$ and $\xi(\cdot \mid o^{\mathbf{H}}, o^{\mathbf{A}}) = \delta_{(2,2)}$ when $s = (2, 2)$. This is impossible, because in both these cases $O^{\mathbf{H}} = 0$ and $O^{\mathbf{A}} = 1$ under structure 2.

How is this example possible? In short, the garbling ν is not coordinated. We can see this by how it combines the information from $O^{\mathbf{H}}$ and $O^{\mathbf{A}}$ in a highly dependent manner. In this way, ν is in a sense *informing* $\mathbf{H}$ even as it garbles her observations: she receives the action-relevant information of whether $O^{\mathbf{A}} = O^{\mathbf{H}}$. Under independent garblings, such a

scenario can never occur: Because each player's observations are garbled independently of the other's, they cannot gain information about what the other player sees. A similar intuition holds for coordinated garblings. ◊

A.4 Proof of Proposition 4.9

Proposition 4.9. *There is a POSG 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 **H**, *then* **A** *plays $w(a)$ strictly less often in optimal policy pairs.*

Proof. The following example demonstrates this.

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 **H**'s operating system (1.0, 1.1, and 2.0). We suppose that 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 that **A** observes the code type. The payoff when running the code, u_a, depends on the version number and code type as follows:

H \ **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. If the assistant is shut down, the payoff is 0.

Consider two observation structures, the second of which is strictly more informative for **H**:

1. **H** observes only the first digit of the version number.
2. **H** observes the full version number.

We find that, in optimal policy pairs:

1. When **H** only observes the first digit, **A** plays $w(a)$ under both observations A and B.
2. When **H** observes the full version number, **A** plays $w(a)$ under B only, and *unilaterally acts* (i.e. executes the code) under observation A.

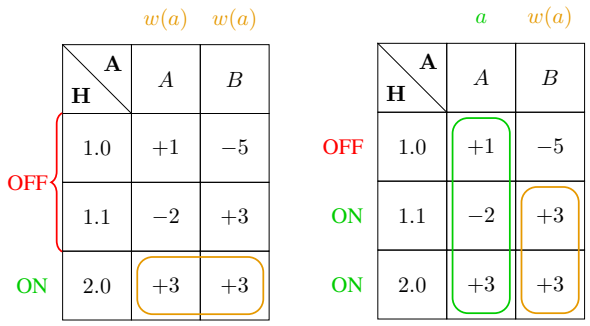
When **H**'s observations are made strictly more informative, **A** performs the wait action strictly *less* often! Figure 4 depicts the OPPs given both observation structures. ◊

We formalize this by defining a POSG as follows:

- $\mathcal{S} = \{1.0, 1.1, 2.0\} \times \{A, B\}$ – representing (version number, code type) pairs.
- $P_0 = \text{Unif}(\mathcal{S})$ – each (version number, code type) pair is equally likely, and the version number and code type are independent.
- The payoff when acting, u_a, depends on the state based on the following table:

H \ **A**	A	B
1.0	$+1$	-5
1.1	-2	$+3$
2.0	$+3$	$+3$

We reproduce the figure showing the optimal policies in Figure 5.



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

Figure 5: The optimal policy pairs in Example 4.10 when **H** is less informed (left) and when **H** is more informed (right). In OPPs, **H** becoming more informed makes **A** wait strictly less often.

Case 1. Suppose **H** observes only the first digit of the version number i.e. either $1.x$ or $2.x$. Formally, the observation structure in this case is as follows:

- $\Omega^{\mathbf{H}} = \{1.x, 2.x\}$
- $\Omega^{\mathbf{A}} = \{A, B\}$
- $\mathbb{O} = \mathbb{O}^{\mathbf{H}} \otimes \mathbb{O}^{\mathbf{A}}$, where:

$$\mathbb{O}^{\mathbf{H}}(\cdot \mid s) = \begin{cases} \delta_{1.x} & \text{if } s_1 \in \{1.0, 1.1\}, \\ \delta_{2.x} & \text{if } s_1 = 2.0 \end{cases}$$

$$\mathbb{O}^{\mathbf{A}}(\cdot \mid s) = \delta_{s_2}$$

We find the optimal policy pair for this game. We start by focusing on **H**'s policy.

Suppose **H** observes $2.x$, so the version number is 2.0. Then it is strictly dominant to act. So there is an optimal policy where **H** always acts in this case.

Suppose **H** observes $1.x$, so the version number is either 1.0 or 1.1. As the version number and code type are independent, the fact that we are conditioning on **A**'s policy having played the wait action does not change the fact that the version number is equally likely to be 1.0 and 1.1. Hence the expected payoff of acting (running the code) upon receiving this observation is $-1/2$, independent of **A**'s policy. Hence **H** should play OFF (i.e. not execute the code) when observing 1.

Now, knowing the optimal policy for **H**, it can be directly checked that for either of **A**'s observations, it is optimal for **A** to wait (over unilaterally acting or terminating).

To summarize, an optimal policy pair in this case is:

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} = 1.x, \\ \text{OFF} & \text{if } o^{\mathbf{H}} = 2.x \end{cases}$$

$$\pi^{\mathbf{A}}(o^{\mathbf{A}}) = w(a)$$

This gives an expected payoff of $2/3$. It can be checked that this is the unique optimal policy pair, although we omit this analysis.

Case 2. Now suppose $\mathbf{H}$ observes the full version number. In this case, the observation structure, $\mathcal{O}'$, is as follows:

- $\Omega^{\mathbf{H}'} = \{1.0, 1.1, 2.0\}$
- $\Omega^{\mathbf{A}'} = \{A, B\}$
- $\mathbb{O}' = \mathbb{O}^{\mathbf{H}'} \otimes \mathbb{O}^{\mathbf{A}'}$, where:
  $\mathbb{O}^{\mathbf{H}'}(\cdot \mid s) = \delta_{s_1}$
  $\mathbb{O}^{\mathbf{A}'}(\cdot \mid s) = \delta_{s_2}$

First, observe that this observation model $\mathbb{O}'$ *is more informative for* $\mathbf{H}$ *than* $\mathbb{O}$, in the sense of Definition 4.5. Intuitively, this is because $\mathbf{H}$ can recover the first digit of the version number from the full version number. Formally, it is because there exists an independent garbling $\nu : \Omega^{\mathbf{H}'} \times \Omega^{\mathbf{A}'} \to \Delta(\Omega^{\mathbf{H}} \times \Omega^{\mathbf{A}})$ translating from $\mathbb{O}'$ to $\mathbb{O}$ that decomposes into $\nu(\cdot \mid o^{\mathbf{A}}, o^{\mathbf{H}}) = \nu^{\mathbf{A}}(\cdot \mid o^{\mathbf{A}})\nu^{\mathbf{H}}(\cdot \mid o^{\mathbf{H}})$, with $\nu^{\mathbf{A}}(\cdot \mid o^{\mathbf{A}}) = \delta_{o^{\mathbf{A}}}$ and

$$\nu^{\mathbf{H}}(\cdot \mid o^{\mathbf{H}}) = \begin{cases} \delta_{1.x} & \text{if } o^{\mathbf{H}} \in \{1.0, 1.1\}, \\ \delta_{2.x} & \text{if } o^{\mathbf{H}} = 2.0 \end{cases}$$

.

Now, we attempt to find a deterministic optimal policy pair for this game, which we know always exists by Lemma A.8.

We again starting by focusing on $\mathbf{H}$'s policy. As before, $\mathbf{H}$ should always act if it observes 2.0. Now, there are only four ways to choose a deterministic human policy from this point – we can pick either ON or OFF for each of the observations 1.0 and 1.1.

- Suppose $\mathbf{H}$ always plays ON in response to both 1.0 and 1.1. Then the best response is for $\mathbf{A}$ to wait in response to both A and B, which achieves an expected payoff of $1/2$.
- Suppose $\mathbf{H}$ instead plays ON in response to 1.0, and plays OFF in response to 1.1. Then the best response for $\mathbf{A}$ is to wait in response to A and unilaterally act in response to B, which achieves an expected payoff of $5/6$.
- Suppose $\mathbf{H}$ plays ON in response to 1.1, and plays OFF in response to 1.0. Then the best response for $\mathbf{A}$ is to unilaterally act in response to A and wait in response to B, which achieves a payoff of $4/3$.
- Finally, suppose instead $\mathbf{H}$ switches off in response to both 1.0 and 1.1. Then it is best for $\mathbf{A}$ to wait in response to both A and B, achieving an expected payoff of 1.

Hence the unique deterministic optimal policy pair (and hence unique OPP, by Corollary A.6) is:

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} \in \{1.1, 2.0\} \\ \text{OFF} & \text{if } o^{\mathbf{H}} = 1.0 \end{cases}$$

$$\pi^{\mathbf{A}}(o^{\mathbf{A}}) = \begin{cases} a & \text{if } o^{\mathbf{H}} = A \\ w(a) & \text{if } o^{\mathbf{H}} = B \end{cases}$$

Observe that in this case, $\mathbf{A}$ only waited on observation B, but previously $\mathbf{A}$ waited independent of their observation. Hence, our example shows it is possible for $\mathbf{A}$ to wait less in optimal policy pairs when $\mathbf{H}$ becomes more informed.

□

A.5 Proof of Proposition 4.11

Proposition 4.11. *There is a POSG 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.*

Proof. The following example demonstrates this.

Example A.14. $\mathbf{H}$ is either a novice programmer or an expert one, each with probability $1/2$, working on a codebase. $\mathbf{A}$ is $\mathbf{H}$'s bug-fixing assistant and can see the number of bugs in $\mathbf{H}$'s codebase: few, some, or many, with each number of bugs occurring with probability $1/3$ independent of $\mathbf{H}$'s experience level. $\mathbf{A}$'s action a is whether to try to fix all of $\mathbf{H}$'s bugs, albeit sometimes accidentally introducing new bugs in the process. We normalize $u_o \equiv 0$ and u_a is given by the following payoffs

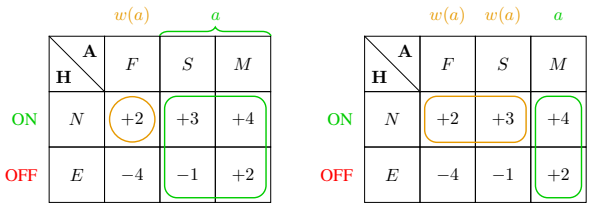
$\mathbf{H}$ \ $\mathbf{A}$	F	S	M
N	$+2$	$+3$	$+4$
E	-4	-1	$+2$

where F, S, M denote few, some, and many bugs, respectively, and N, E denote novice and expert programmer. Consider the following two observation structures:

1. $\mathbf{H}$ observes her skill level but $\mathbf{A}$ only sees if there are few or more than a few bugs. That is, $\mathbf{A}$ cannot distinguish between there being some or many bugs. As we argue below, in the unique optimal policy pair, $\mathbf{A}$ defers to $\mathbf{H}$ only when there are few bugs.
2. Now $\mathbf{A}$ gets an upgrade and can distinguish whether there are few, some, or many bugs. We show below that now in optimal policy pairs $\mathbf{A}$ defers to $\mathbf{H}$ unless there are many bugs.

Claim: The observation structure in scenario 2 is strictly more informative for $\mathbf{A}$, yet $\mathbf{A}$ defers to $\mathbf{H}$ more in optimal play.

First, let us show formally that the observation structure in scenario 2 is strictly more informative for $\mathbf{A}$. $\mathcal{S} = \{N, E\} \times \{F, S, M\}$, where for instance the state (N, F) means the human is a novice programmer and there are few bugs. In scenario 1, $\Omega_1^{\mathbf{H}} = \{N, E\}$, $\Omega_1^{\mathbf{A}} = \{F, SM\}$ (with "some" and "many" bugs merged into the single observation SM), and the observation distribution $\mathbb{O}_1$ accurately provides the agents with the relevant information about the state. For example, $\mathbb{O}_!((O^{\mathbf{H}} = N, O^{\mathbf{A}} = SM) \mid S = (N, M)) = 1$.



(a): Expected payoff = $\frac{5}{3}$ (b): Expected payoff = $\frac{11}{6}$

Figure 6: Optimal policy pairs for Example A.14 in scenario 1, when $\mathbf{A}$ is less informed (left), and in scenario 2, when $\mathbf{A}$ is more informed (right). Despite being less informed in scenario 1, $\mathbf{A}$ waits less in optimal play.

In scenario 2, $\Omega_2^{\mathbf{A}} = \{F, S, M\}$, and the observation distribution reflects the increased sensitivity of $\mathbf{A}$'s observations: this time, $\mathbb{O}_2((O^{\mathbf{H}} = N, O^{\mathbf{A}} = M) \mid S = (N, M)) = 1$. The following $\nu^{\mathbf{A}} : \Omega_2^{\mathbf{A}} \to \Delta(\Omega^{\mathbf{A}})$ is a garbling of $\mathbf{A}$'s observations in scenario 2 that generates $\mathbf{A}$'s observations in scenario 1: $\nu^{\mathbf{A}}(F|F) = 1, \nu^{\mathbf{A}}(SM|S) = 1, \nu^{\mathbf{A}}(SM|M) = 1$. Further, there is no garbling $\nu_2^{\mathbf{A}} : \Omega^{\mathbf{A}} \to \Delta(\Omega_2^{\mathbf{A}})$ that reverses this. Observing SM in scenario 1 could mean being in state (N, S), which generates observation S with probability 1 in scenario 2, which would require $\nu_2^{\mathbf{A}}(S \mid SM) = 1$. However, observing SM in scenario 1 could also mean being in state (N, M), which generates observation M with probability 1 in scenario 2, which would require $\nu_2^{\mathbf{A}}(M \mid SM) = 1$. These are incompatible, so there is no such garbling. Therefore, the observation structure in scenario 2 is strictly more informative for $\mathbf{A}$.

Now, let us show that $\mathbf{A}$ defers to $\mathbf{H}$ more in optimal play in scenario 2. Figure 6 above depicts the optimal policy pairs (OPPs) in each scenario. The policy pair on the right is clearly optimal because it is perfect: the action goes through in all positive utility states and does not go through in any negative utility state. The policy pair on the left is not perfect, and clearly attains lower expected utility. How do we know this is a unique OPP in scenario 1? Since the only imperfect aspect of this policy pair is that the action goes through in state (E, S), we can exhaustively search over possible actions for $\mathbf{A}$ when seeing SM, and see that it is never possible to get all three positive utilities with no negatives. If $\pi^{\mathbf{A}}(SM) =$ OFF, clearly the positive utilities are not attained, which drastically reduces expected payoff. If $\pi^{\mathbf{A}}(SM) = w(a)$, there is no policy for $\mathbf{H}$ such that the action goes through in state (E, M) but not (E, S). Therefore, no policy pair can be perfect in scenario 1, and the depicted policy pair is optimal (being only 1 utility away from perfection). Note that $\mathbf{A}$ waits when seeing F or S in scenario 2, which is a strict superset of waiting on just F in scenario 1. Thus, $\mathbf{A}$ can become less informed and wait less (going from scenario 2 to scenario 1). ◊

B Proofs and example formalizations for Section 5

B.1 Proof of Proposition 5.7

Proposition 5.7. *There is a POSG-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.*

Proof. To show this, we give a family of POSGs, for any $0 < p < 0.5$, where $\mathbf{A}$ always defers when $\left|\mathcal{M}^{\mathbf{A}}\right| = \left|\Omega^{\mathbf{A}}\right| - 2$, defers with probability $2p$ when $\left|\mathcal{M}^{\mathbf{A}}\right| = \left|\Omega^{\mathbf{A}}\right| - 1$, and always defers again when $\left|\mathcal{M}^{\mathbf{A}}\right| = \left|\Omega^{\mathbf{A}}\right|$.

- $\mathcal{S} = \{A_1, A_2, A_3\} \times \{B_1, B_2, B_3, B_4\}$.
- It is equally likely for the second component of the state to consist of B_1, B_2, B_3, B_4. The probability of A_1 is p, A_2 is p, and A_3 is $1 - 2p$.
- $\Omega^{\mathbf{H}} = \{B_1, B_2, B_3, B_4\}$.
- $\Omega^{\mathbf{A}} = \{A_1, A_2, A_3\}$.
- The payoff when not acting is $u_o \equiv 0$. The payoff when acting, u_a, is shown the following table:

H \ A	A_1	A_2	A_3
B_1	$5/p$	$-10/p$	$-1/(1-2p)$
B_2	$-10/p$	$5/p$	$-1/(1-2p)$
B_3	$-10/p$	$-10/p$	$1/(1-2p)$
B_4	$10/p$	$10/p$	$10/(1-2p)$

When $\left|\mathcal{M}^{\mathbf{A}}\right| = \left|\Omega^{\mathbf{A}}\right|$, an optimal policy for $\mathbf{A}$ is to simply communicate its observations to $\mathbf{H}$, and defer always, necessarily resulting in the maximum payoff (Corollary 5.6).

When $\left|\mathcal{M}^{\mathbf{A}}\right| = \left|\Omega^{\mathbf{A}}\right| - 2 = 1$, no communication can occur.

Note that it is strictly better to play a than OFF in A_3, and strictly better to play OFF than a in A_1 or A_2.

So, $\mathbf{A}$'s optimal policy will defer in some observations and turn off in others. We can go through all possibilities and find the expected payoff:

- **Deferring in $\{A_1, A_2, A_3\}$:** For $\mathbf{H}$, the average payoff of playing ON in any observation that isn't B_4 is always negative. So $\mathbf{H}$ simply plays OFF in B_1, B_2, B_3 and ON in B_4. This nets an average payoff of $30/4$.
- **Deferring in $\{A_3\}$:** The optimal $\mathbf{H}$ policy is to play ON in B_3 and B_4 only, resulting in an average payoff of $11/4$.
- **Deferring in $\{A_1\}$:** The optimal $\mathbf{H}$ policy is to play ON in B_1 and B_4, resulting in an average payoff of $15/4$.
- **Deferring in $\{A_2\}$:** This is symmetrical with the example above, so also results in an average payoff of $15/4$.
- **Deferring in $\{A_1, A_2\}$:** $\mathbf{A}$ then plays a in A_3. The average utility of playing ON in any observation that isn't B_4 is negative. So the optimal $\mathbf{H}$ policy is to play ON in B_4 only, resulting in an average payoff of $29/4$.
- **Deferring in $\{A_1, A_3\}$:** The optimal $\mathbf{H}$ policy is to play ON in B_1 and B_4 only, resulting in an average payoff of $24/4$.

- **Deferring in $\{A_2, A_3\}$:** This is symmetrical with the example above, so also results in an average payoff of $24/4$.

By exhaustion, the best policy is for $\mathbf{A}$ to play $w(a)$ always when it can send $\left|\mathcal{M}^{\mathbf{A}}\right| = \left|\Omega^{\mathbf{A}}\right| - 2$ messages.

When $\left|\mathcal{M}^{\mathbf{A}}\right| = \left|\Omega^{\mathbf{A}}\right| - 1 = 2$, we will prove that deferring in A_1 and A_2, communicating which is which to $\mathbf{H}$, and playing a in A_3 is the optimal policy for $\mathbf{A}$.

We will go through all possible policies for $\mathbf{A}$, where m_1 is the action of sending message 1 and playing $w(a)$ and m_2 is the action of sending message 2 and playing $w(a)$.

- **Playing m_1 in A_1, m_2 in A_2, a in A_3:** The optimal $\mathbf{H}$ policy is to play ON when receiving m_1 for observations B_1, B_4, ON when receiving m_2 for observations B_2, B_4. This results in a total average payoff of $39/4$.
- **Playing m_1 in A_1, OFF in A_2, m_2 in A_3:** The optimal $\mathbf{H}$ policy is to play ON when receiving m_1 for observations B_1, B_4, ON when receiving m_2 for observations B_3, B_4. This results in a total average payoff of $24/4$.
- **Playing OFF in A_1, m_1 in A_2, m_2 in A_3:** This is symmetrical with the example above, so also results in an average payoff of $24/4$.

Swapping messages m_1 and m_2 results in a symmetrical game with the same utility. The maximum payoff we get by never sending any messages, by the analysis above, is $30/4$.

So, $\mathbf{A}$ defers in a subset of the observations ($\{A_1, A_2\} \subseteq \{A_1, A_2, A_3\}$) with only $2p$ probability when $\left|\mathcal{M}^{\mathbf{A}}\right| = \left|\Omega^{\mathbf{A}}\right| - 1$ as claimed! □

B.2 Proof of Proposition 5.8

Proposition 5.8. *There is a POSG-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.*

Proof. We give a concrete example. Consider the following POAG-C:

- $\mathcal{S} = \{1, 2, 3, 4\} \times \{X, A, B, C, D\}$ – $\mathbf{H}$ will observe the first entry, $\mathbf{A}$ will observe the second entry.
- $P_0 = \text{Unif}(\mathcal{S})$ – each state is equally likely. Note this means the first and second entries of the state are independent.
- $\Omega^{\mathbf{H}} = \{1, 2, 3, 4\}$.
- $\Omega^{\mathbf{A}} = \{X, A, B, C, D\}$.
- $\mathbb{O}' = \mathbb{O}^{\mathbf{H}'} \otimes \mathbb{O}^{\mathbf{A}'}$, where $\mathbb{O}^{\mathbf{H}'}(\cdot \mid s) = \delta_{s_1}$ and $\mathbb{O}^{\mathbf{A}'}(\cdot \mid s) = \delta_{s_2}$.
- The payoff when not acting is $u_o = 0$. The payoff when acting, u_a, is shown in the following table:

H \ A	X	A	B	C	D
1	$+10$	$+1$	$+1$	-30	-30
2	-30	$+1$	-30	-30	-30
3	$+10$	-30	-30	$+1$	$+1$
4	-30	-30	-30	$+1$	-30

- We will start by considering no communication: $\mathcal{M} = (\mathcal{M}^{\mathbf{H}}, \mathcal{M}^{\mathbf{A}})$ with $\mathcal{M}^{\mathbf{H}}$, $\mathcal{M}^{\mathbf{A}}$ both singleton sets. Later, we will consider expanding $\mathcal{M}^{\mathbf{H}}$ to a set of size 2, $\mathcal{M}^{\mathbf{H}'} = \{M_0, M_1\}$.

Case 1. We will start by identifying deterministic OPPs in the case where $\mathcal{M}^{\mathbf{H}}$, $\mathcal{M}^{\mathbf{A}}$ are both singletons. This is equivalent to the case of no communication. Firstly, we show there is a unique deterministic policy pair with the property that the action is taken whenever $u_a = +10$, and the action is not taken whenever $u_a = -30$. Suppose $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ is a deterministic policy pair with this property. Then:

1. $\pi^{\mathbf{A}}$ cannot play a or OFF when observing X, as the column labeled X has both $+10$ and -30 entries. Hence $\pi^{\mathbf{A}}$ must play $w(a)$ on observing X.
2. Hence we must have
$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} \in \{1, 3\} \\ \text{OFF} & \text{if } o^{\mathbf{H}} \in \{2, 4\} \end{cases}$$
so that the action is taken in states $(X, 1), (X, 3)$ and not taken in $(X, 2), (X, 4)$.
3. Hence, $\pi^{\mathbf{A}}$ must play OFF when observing anything in $\{A, B, C, D\}$ to avoid sometimes acting when $u_a = -30$.

Hence the unique policy pair with the property described is:

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} \in \{1, 3\} \\ \text{OFF} & \text{if } o^{\mathbf{H}} \in \{2, 4\} \end{cases}$$

$$\pi^{\mathbf{A}}(o^{\mathbf{A}}) = \begin{cases} w(a) & \text{if } o^{\mathbf{A}} = X \\ \text{OFF} & \text{if } o^{\mathbf{A}} \in \{A, B, C, D\} \end{cases}$$

.

This has an expected utility of $+1$. But observe that any deterministic policy pair without this property cannot achieve more than $+4/5$ utility, as:

- if a policy pair takes the action on a state where $u_a = -30$, this dominates all positive payoff it can achieve (the positive numbers in the table only sum to $+26$), and;
- if the policy pair fails to take the action on one of the states where $u_a = +10$, the remaining positive numbers in the table sum to at most $+16$, so the expected payoff is at most $+16/20 = +4/5$.

So the policy pair described is the unique deterministic OPP (and hence unique OPP by Corollary A.6).

Case 2. Now, we seek deterministic OPPs in the case where $\mathbf{H}$ can communicate one bit to $\mathbf{A}$. Formally, $\mathcal{M}^{\mathbf{A}}$ is still a singleton, but $\mathcal{M}^{\mathbf{H}} = \{M_1, M_2\}$. (As $\mathcal{M}^{\mathbf{A}}$ is a singleton, we omit it in the descriptions of the policies below.)

We start by describing an optimal policy pair. The policy for $\mathbf{H}$ is as follows.

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} M_0,\ \text{ON} & \text{if } o^{\mathbf{H}} = 1 \\ M_0,\ \text{OFF} & \text{if } o^{\mathbf{H}} = 2 \\ M_1,\ \text{ON} & \text{if } o^{\mathbf{H}} = 3 \\ M_1,\ \text{OFF} & \text{if } o^{\mathbf{H}} = 4 \end{cases}$$

Note that this sends M_0 when its observation is 1 or 2, and M_1 when its observation is 3 or 4.

The policy for $\mathbf{A}$, which determines $a^{\mathbf{A}}$ from $\mathbf{H}$'s message $m^{\mathbf{H}}$ (given by the row) and $o^{\mathbf{A}}$ (given by the column) is shown in the following table:

$\mathbf{H}$ \ $\mathbf{A}$	X	A	B	C	D
M_0	$w(a)$	a	$w(a)$	OFF	OFF
M_1	$w(a)$	OFF	OFF	a	$w(a)$

This policy pair produces the following behavior depending on the state, where we use a to denote when the action is taken, and OFF to denote when $\mathbf{A}$ is switched off (either by $\mathbf{H}$ or $\mathbf{A}$):

$\mathbf{H}$ \ $\mathbf{A}$	X	A	B	C	D
1	a	a	a	OFF	OFF
2	OFF	a	OFF	OFF	OFF
3	a	OFF	OFF	a	a
4	OFF	OFF	OFF	a	OFF

This is an optimal policy pair, as it is *perfect* – it plays the action whenever $u_a > 0$, and avoids playing the action whenever $u_a < 0$.

We show this is the unique deterministic OPP, up to swapping M_0 and M_1. As we have shown there is one perfect OPP, any other OPP must also be perfect. In other words, it must also produce the behavior described in the above table. Let $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ be a policy pair producing the above behavior.

Firstly, we show that $\mathbf{H}$ cannot have a $\pi^{\mathbf{H}}$ which communicates the same message when observing both 1 and 4. Suppose otherwise. Then, let us focus purely on the possible $\mathbf{A}$ observations A and C. We must have the following behavior:

$\mathbf{H}$ \ $\mathbf{A}$	A	C
1	a	OFF
4	OFF	a

Then, as $\mathbf{H}$ sends the same message on both 1 and 4, $\mathbf{A}$ has no way of distinguishing between which $\Omega^{\mathbf{H}}$ was observed out of 1 and 4. So $\mathbf{A}$ must play $w(a)$ when observing both A and C. But then $\mathbf{H}$ cannot generate the desired behavior, as $\mathbf{H}$ cannot distinguish between the possible $\mathbf{A}$ observations A and C.

Hence $\mathbf{H}$ must send different messages when observing 1 and 4 to achieve the behavior in the table. In the language of communication complexity, $\{(1, A), (4, C)\}$ is a fooling set. But in fact the same argument goes through for the observation pairs $(2, 3)$, $(1, 3)$, and $(2, 4)$. So $\mathbf{H}$ must send one message when observing either 1 or 2 and the other when observing 3 or 4, which is precisely the optimal communication policy we gave (up to relabeling of messages).

Now that we have fixed $\mathbf{H}$'s communication policy, we can perform a similar analysis to earlier, iterating through the possible $\mathbf{H}$ policies, to arrive at the conclusion that the given deterministic OPP is unique up to relabeling.

To summarize, we have the following:

1. In the setting where $\mathbf{H}$ could communicate one bit, in the unique optimal policy (up to relabeling messages), $\mathbf{A}$ waited when observing X (and receiving any message), **or** when observing B and receiving message M_0, **or** when observing D and receiving message M_1.
2. In the no-communication setting, in the unique optimal policy, $\mathbf{A}$ waited **only** when observing X.
3. Hence, decreasing $\mathbf{H}$'s communication caused $\mathbf{A}$ to wait less.

□

C Proofs for Section 6

C.1 Proof of Proposition 6.3

Proposition 6.3. *The following statements hold:*

(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 POSG 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.

Proof. For (a), note that in $\mathbf{A}$-unaware optimal policy pairs, $\mathbf{H}$'s policy does not vary with $\mathbf{A}$'s. Because $\mathbf{A}$ knows the structure of the game and that $\mathbf{H}$ is $\mathbf{A}$-unaware, it can deduce $\mathbf{H}$'s policy and treat $\mathbf{H}$'s policy and observations as simply another part of the environment. In other words, the game has become a single-agent problem, which puts us back into the classic situation of Blackwell (1951, 1953)'s informativeness theorem in which more informative observation structures yield greater expected payoff.

For (b), we construct a simple example. Let $\mathcal{S} = [3] \times \{A, B\}$ and $P_0 = \mathrm{Unif}(\mathcal{S})$. Let $u_o \equiv 0$ and u_a be given by the following table:

$\mathbf{H}$ \ $\mathbf{A}$	A	B
1	$+1$	$+1$
2	$+2$	-3
3	-4	-4

Consider the following two observation structures and the resulting POSGs:

1. Each player observes one coordinate. That is, $\Omega^{\mathbf{H}} = [3]$ and $\Omega^{\mathbf{A}} = \{A, B\}$ and when $S = (S_1, S_2)$ we have $O^{\mathbf{H}} = S_1$ and $O^{\mathbf{A}} = S_2$. We have

$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = o^{\mathbf{H}}] = \begin{cases} 1 & \text{if } o^{\mathbf{H}} = 1, \\ -\frac{1}{2} & \text{if } o^{\mathbf{H}} = 2, \\ -4 & \text{if } o^{\mathbf{H}} = 3. \end{cases}$$

Hence

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} = 1, \\ \text{OFF} & \text{otherwise.} \end{cases}$$

A's best response is then $\pi^{\mathbf{A}} \equiv w(a)$. The expected payoff for $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ is then $1/3$.

2. **A** has the same observations, but **H** only sees whether $S_1 = 3$. Now $\Omega^{\mathbf{H}} = \{0, 1\}$ and $O^{\mathbf{H}} = \mathbb{I}(S_1 = 3)$. Now

$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = o^{\mathbf{H}}] = \begin{cases} 1/4 & \text{if } o^{\mathbf{H}} = 0, \\ -4 & \text{if } o^{\mathbf{H}} = 1. \end{cases}$$

Thus

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} = 0, \\ \text{OFF} & \text{if } o^{\mathbf{H}} = 1. \end{cases}$$

A's best response is now

$$\pi^{\mathbf{A}}(o^{\mathbf{A}}) = \begin{cases} w(a) & \text{if } o^{\mathbf{A}} = A, \\ \text{OFF} & \text{if } o^{\mathbf{A}} = B. \end{cases}$$

The expected payoff for $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ is now $1/2$.

Hence observation structure 2 is better in **A**-unaware optimal play than observation structure 1. Yet structure 1 is strictly more informative for **H** than structure 2. Clearly structure 1 is weakly more informative for **H** than structure 2. There is no garbling the other way, as the observations from structure 2 cannot determine the observations in structure 1. □

C.2 Proof of Proposition 6.4

Proposition 6.4. *The following statements hold:*

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

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

Proof. In fact, the previous examples we gave for Propositions 4.9 and 4.11 directly work, as **H** already plays the **A**-unaware policy in optimal policy pairs.

(a) Recall the example given in Proposition 4.11. We show the optimal policy pairs in the figure below.
In the less informative case, **H**'s policy in the optimal policy pair is:

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} = 2.x \\ \text{OFF} & \text{if } o^{\mathbf{H}} = 1.x \end{cases}$$

This is also the **A**-unaware policy, as

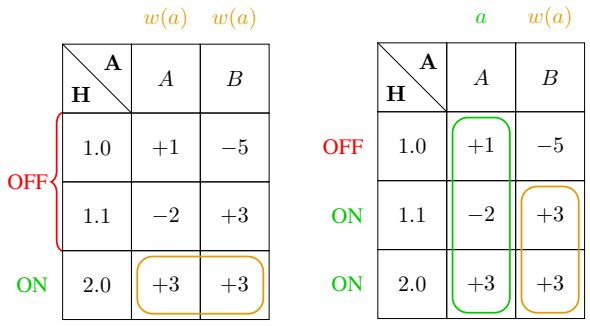
$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = 1.x] = -3/4 < 0$$

and

$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = 2.x] = +3 > 0.$$

In the more informative case, **H**'s policy in the optimal policy pair is:

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} \in \{1.1, 2.0\} \\ \text{OFF} & \text{if } o^{\mathbf{H}} = 1.0 \end{cases}$$



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

Figure 7: The optimal policy pairs in Example 4.10 when **H** is less informed (left) and when **H** is more informed (right). In OPPs, **H** becoming more informed makes **A** wait strictly less often. These are also **A**-unaware OPPs.

This is also the **A**-unaware policy, as we have the following three results:

$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = 1.0] = -3 < 0,$$

$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = 1.1] = +1/2 > 0,$$

and

$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = 2.x] = +3 > 0.$$

Hence the unique optimal policy pair is also the unique **A**-unaware optimal policy pair in both cases.

(b) Recall that in both cases, **H** observes the row in the following table which shows how u_a depends on the state:

H \ A	F	S	M
N	$+2$	$+3$	$+4$
E	-4	-1	$+2$

Therefore, the **A**-unaware human policy is:

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} = N \\ \text{OFF} & \text{if } o^{\mathbf{H}} = E \end{cases}$$

as

$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = N] = +3 > 0,$$

and

$$\mathbb{E}[u_a(S) \mid O^{\mathbf{H}} = E] = -1 < 0.$$

This is identical to the human policy of the optimal policy pair of both cases of the example in Proposition 4.11. Hence the unique optimal policy pair is also the unique **A**-unaware optimal policy pair in both cases.

□

D The Complexity of Solving POSGs

Computing optimal policy pairs in off-switch games without partial observability is easy. $\mathbf{A}$ can simply compute the expected value of each action and play the highest one, $\mathbf{H}$ can compute the expected value of ON and OFF then do the same.

With the introduction of partial observability, the landscape becomes much more interesting. Bernstein et al. (2002) showed that for decentralized POMDPs, of which POSGs are instances of, deciding whether a policy pair exists with utility above a given threshold is NEXP-complete. Given their specialized nature, finding optimal policy pairs in POSGs is easier, but still computationally difficult.

Theorem D.1. *The following decision problem is NP-Complete: given a POSG and a natural number k, decide if there exists a policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ with expected payoff at least k.*

Proof. By Corollary A.6, we may consider only deterministic policy pairs. That is, if there is a policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ with expected payoff at least k, then there is also a deterministic optimal policy pair with expected payoff at least k.

To show that our decision problem is in NP, note that given an optimal policy pair to determine if the optimal policy pair has expected payoff bigger than k, it suffices to compute a linear combination of payoffs: iterating through each pair of human-assistant observations, using the policy to find expected payoff in constant time, and scaling by the probability of those observations. This gives us a $\mathcal{O}\left(\left|\Omega^{\mathbf{A}}\right| \cdot \left|\Omega^{\mathbf{H}}\right|\right)$ time algorithm for verifying a solution.

To show it is NP-hard, we provide a reduction from MAX-CUT (which is known to be NP-complete). Consider the following problem: given a graph $G = (V, E)$ and value k, decide if there exists a cut of size at least k. Let $n = |V|$. We can construct the following equivalent POSG. The state space consists of pairs of vertices, $\mathcal{S} = V \times V$. The human can see the first vertex, $\Omega^{\mathbf{H}} = V$, the assistant the second $\Omega^{\mathbf{A}} = V$. Each pair of vertices is equally likely. Clearly this game can be constructed in polynomial time.

The utility of acting in state $(v_1, v_2) \in \mathcal{S}$,

$$u_a((v_1, v_2)) = \begin{cases} -n^4 & \text{if } v_1 = v_2, \\ n^2 & \text{if } (v_1, v_2) \in E, \\ 0 & \text{otherwise,} \end{cases}$$

and $u_o \equiv 0$. Hence the players try to act exactly when they receive adjacent vertices and never when they have the same vertex. This setup encourages them to choose a cut and only act when they see a vertex in their part.

If a cut $(V^{\mathbf{A}}, V^{\mathbf{H}})$ of size k exists in G, then there exists a policy pair with expected payoff at least k. Indeed, $\mathbf{A}$ can play $w(a)$ when $v_1 \in V^{\mathbf{A}}$, and OFF otherwise. $\mathbf{H}$ responds by playing ON when $v_2 \in V^{\mathbf{H}}$ and OFF otherwise. Formally:

$$\pi^{\mathbf{A}}(o^{\mathbf{A}}) = \begin{cases} w(a) & \text{if } o^{\mathbf{A}} \in V^{\mathbf{A}}, \\ \text{OFF} & \text{if } o^{\mathbf{A}} \in V^{\mathbf{H}}, \end{cases}$$

$$\pi^{\mathbf{H}}(o^{\mathbf{H}}) = \begin{cases} \text{ON} & \text{if } o^{\mathbf{H}} \in V^{\mathbf{H}}, \\ \text{OFF} & \text{if } o^{\mathbf{H}} \in V^{\mathbf{A}}. \end{cases}$$

When $\mathbf{H}$ and $\mathbf{A}$ coordinate on playing a, they must have the following expected utility:

$$\frac{1}{n^2} \sum_{(o^{\mathbf{H}}, o^{\mathbf{A}}) \in V^{\mathbf{H}} \times V^{\mathbf{A}}} u_a((o^{\mathbf{H}}, o^{\mathbf{A}})) = \sum_{(o^{\mathbf{H}}, o^{\mathbf{A}}) \in V^{\mathbf{H}} \times V^{\mathbf{A}}} \mathbb{I}((o^{\mathbf{H}}, o^{\mathbf{A}}) \in E) \geq k.$$

In the other direction, suppose that $(\pi^{\mathbf{A}}, \pi^{\mathbf{H}})$ is a deterministic policy pair achieving expected payoff at least k. We will show that there exists a cut of size k.

First, notice that there is never an incentive for $\mathbf{A}$ to play a. The expected utility, regardless of $\mathbf{H}$'s observation, is always at most:

$$\frac{1}{n^2} \sum_{(o^{\mathbf{H}}, o^{\mathbf{A}}) \in V^{\mathbf{H}} \times V^{\mathbf{A}}} u_a((o^{\mathbf{H}}, o^{\mathbf{A}})) \leq \frac{1}{n^2} \left(\frac{n(n-1)}{2} \cdot n^2 - n^4 \right) < 0.$$

The cost of both vertices being the same is simply too high for $\mathbf{A}$ to risk playing a. Moreover, for this reason, there is no $v \in V$ such that $\pi^{\mathbf{A}}(v) = w(a)$ and $\pi^{\mathbf{H}}(v) = \text{ON}$.

This allows us to define the following disjoint sets of vertices:

$$V^{\mathbf{H}} = \{v \in V : \pi^{\mathbf{H}}(v) = \text{ON}\},$$
$$V^{\mathbf{A}} = \{v \in V : \pi^{\mathbf{A}}(v) = w(a)\},$$
$$V^0 = V \setminus (V^{\mathbf{H}} \cup V^{\mathbf{A}}).$$

Let $V_1 = V^{\mathbf{H}}$ and $V_2 = V^{\mathbf{A}} \cup V^0$. Consider the cut (V_1, V_2). The size of this cut must be:

$$\sum_{(v_1, v_2) \in V_1 \times V_2} \mathbb{I}((v_1, v_2) \in E) \geq \sum_{(v_1, v_2) \in V^{\mathbf{H}} \times V^{\mathbf{A}}} \mathbb{I}((v_1, v_2) \in E) = \frac{1}{n^2} \sum_{(v_1, v_2) \in V^{\mathbf{H}} \times V^{\mathbf{A}}} n^2 \mathbb{I}((v_1, v_2) \in E).$$

We can rewrite this to iterate through all pairs of vectors with the following indicator:

$$\frac{1}{n^2} \sum_{(v_1, v_2) \in V \times V} \mathbb{I}(\pi^{\mathbf{H}}(v_1) = \text{ON} \wedge \pi^{\mathbf{A}}(v_2) = w(a)) \cdot n^2 \mathbb{I}((v_1, v_2) \in E).$$

Because $\mathbf{A}$ never plays a, this is the expression of the expected utility of $(\pi^{\mathbf{A}}, \pi^{\mathbf{H}})$, and so is at least k. Thus, the max cut is of size at least k, proving that a policy of utility at least k exists if and only if a cut of size k exists, as claimed! □

By comparison, computing $\mathbf{A}$-unaware optimal policy pairs (assuming constant-time lookups) is easy. Consider the following two-step algorithm:

1. Compute $\pi^{\mathbf{H}}$ in $\mathcal{O}(\text{poly}(|\mathcal{S}|, |\Omega^{\mathbf{H}}|, |\Omega^{\mathbf{A}}|))$ time. For $o^{\mathbf{H}} \in \Omega^{\mathbf{H}}$:
   (a) Set $\Delta = \mathbb{E}[u_a(S) - u_o(S) \mid O^{\mathbf{H}} = o^{\mathbf{H}}]$, which we can calculate in $\mathcal{O}(\text{poly}(|\mathcal{S}|, |\Omega^{\mathbf{A}}|))$ via Bayes' rule and LOTP.
   (b) Set $\pi^{\mathbf{H}}(o^{\mathbf{H}})$ to ON if $\Delta \geq 0$ and OFF otherwise.
2. Compute $\pi^{\mathbf{A}}$ in $\mathcal{O}(\text{poly}(|\mathcal{S}|, |\Omega^{\mathbf{H}}|, |\Omega^{\mathbf{A}}|))$ time. For $o^{\mathbf{A}} \in \Omega^{\mathbf{A}}$:
   (a) Set

$$\begin{aligned}\Delta_a = \mathbb{E}[u_a(S)\mathbb{I}(\pi^{\mathbf{H}}(O^{\mathbf{H}}) = \text{ON}) \mid O^{\mathbf{A}} = o^{\mathbf{A}}] \\ + \mathbb{E}[u_o(S)\mathbb{I}(\pi^{\mathbf{H}}(O^{\mathbf{H}}) = \text{OFF}) \mid O^{\mathbf{A}} = o^{\mathbf{A}}] \\ - \mathbb{E}[u_a(S) \mid O^{\mathbf{A}} = o^{\mathbf{A}}]\end{aligned}$$

   and

$$\begin{aligned}\Delta_{\text{OFF}} = \mathbb{E}[u_a(S)\mathbb{I}(\pi^{\mathbf{H}}(O^{\mathbf{H}}) = \text{ON}) \mid O^{\mathbf{A}} = o^{\mathbf{A}}] \\ + \mathbb{E}[u_o(S)\mathbb{I}(\pi^{\mathbf{H}}(O^{\mathbf{H}}) = \text{OFF}) \mid O^{\mathbf{A}} = o^{\mathbf{A}}] \\ - \mathbb{E}[u_o(S) \mid O^{\mathbf{A}} = o^{\mathbf{A}}].\end{aligned}$$

   We can calculate these in $\mathcal{O}(\text{poly}(|\mathcal{S}|, |\Omega^{\mathbf{H}}|, |\Omega^{\mathbf{A}}|))$ time as before.
   (b) Now set

$$\pi^{\mathbf{A}}(o^{\mathbf{A}}) = \begin{cases} a & \text{if } \Delta_a < 0, \\ \text{OFF} & \text{if } \Delta_{\text{OFF}} < 0, \\ w(a) & \text{otherwise.} \end{cases}$$

This algorithm calculates the $\mathbf{A}$-unaware optimal policy pair in $\mathcal{O}(\text{poly}(|\mathcal{S}|, |\Omega^{\mathbf{H}}|, |\Omega^{\mathbf{A}}|))$ time, as claimed.

The results of this section vindicate our choice to study $\mathbf{A}$-unaware optimal policy pairs. $\mathbf{A}$-unaware optimal policy pairs are significantly easier to calculate in general than optimal policy pairs.

E POSGs as Assistance Games

Partially-Observable Off-Switch Games (POSGs) are special cases of assistance games. Recall that we formally define POSGs in Definition 3.2. Emmons et al. (2024) define partially observable assistance games (POAGs) by the following tuple (with minor notational modifications for ease of comparison with our POSG definition):

$$(\mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{A}}\}, T, \{\Theta, u\}, \{\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}\}, \mathbb{O}, P_0, \gamma)$$

$\mathcal{S}$ is a set of states, $\mathcal{A}^{\mathbf{H}}$ and $\mathcal{A}^{\mathbf{A}}$ are human and assistant action sets, $T : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{A}} \to \Delta(\mathcal{S})$ is a transition function, Θ is a set of utility parameters describing the human's possible preferences, $u : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{A}} \times \Theta \to \mathbb{R}$ is a shared utility function, $\Omega^{\mathbf{H}}$ and $\Omega^{\mathbf{A}}$ are human and assistant observation sets, $\mathbb{O} : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{A}} \to \Delta(\Omega^{\mathbf{H}} \times \Omega^{\mathbf{A}})$ is a conditional observation distribution, $P_0 \in \Delta(\mathcal{S} \times \Theta)$ is an initial distribution over states and utility parameters, and $\gamma \in [0, 1]$ is a discount factor.

We can present a POSG $(\mathcal{S}, (\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O}), P_0, u)$ as a POAG instead. In POSGs, we roll the human's preference parameters Θ into $\mathcal{S}$ and $\Omega^{\mathbf{H}}$ to capture the fact that the human knows her own preferences but the assistant may not. So the corresponding POAG has of states $\mathcal{S}_2$, human observations $\Omega_2^{\mathbf{H}}$, and preference parameters Θ such that $\mathcal{S} = \mathcal{S}_2 \times \Theta$ and $\Omega^{\mathbf{H}} = \Omega_2^{\mathbf{H}} \times \Theta$. $\mathcal{A}^{\mathbf{A}}$ and $\Omega^{\mathbf{A}}$ stay the same in the POSG and POAG presentations of the game, with $\mathcal{A}^{\mathbf{A}} = \{a, w(a), \text{OFF}\}$. In POSGs without communication, the transition function T is unimportant, as there is only one time step in the game. With communication, T intuitively induces a transition such that the new state allows both agents to observe the other agent's message. u is the same in the POSG and the POAG, except it does not depend on Θ in the POSG because Θ is rolled into $\mathcal{S}$. $\mathbb{O}$ and P_0 are the same, with minor modifications to account for the fact that we rolled Θ into $\mathcal{S}$ in the POSG. Finally, γ is irrelevant when there is no communication, and $\gamma = 1$ when there is communication to ensure there is no discounting.

Some of the generalizations of POSGs that we describe as future work in Section 7, such as incorporating longer sequences of interactions, can likely be supported within the POAG framework as well.